NASA/CP-2011-217069/Part 1

Selected Papers and Presentations Presented at MODSIM World 2010 Conference & Expo

Edited by

Thomas E. Pinelli
Langley Research Center, Hampton, Virginia

March 2011

NASA/TM-2011-217069/Part 1

Selected Papers and Presentations Presented at MODSIM World 2010 Conference & Expo

Edited by

Thomas E. Pinelli
Langley Research Center, Hampton, Virginia

Proceedings of a conference sponsored by the
National Aeronautics and Space Administration
and held in Hampton, Virginia
October 13–15, 2010

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

March 2011

The use of trademarks or names of manufacturers in the report is for accurate reporting and does not constitute an official endorsement, either expressed or implied, of such products or manufacturers by the National Aeronautics and Space Administration.

Available from:

NASA Center for AeroSpace Information
7115 Standard Drive
Hanover, MD 21076-1320
443-757-5802

Preface

Modeling and simulation-based (MODSIM) engineering and science is rapidly becoming an essential scientific methodology for nearly all areas of engineering and many branches of science and for research, development, concept generation, product design and manufacturing, and consumer marketing. Continuing advances in computational science and networking technologies have made MODSIM-based engineering and science a "powerful and ubiquitous tool" for engineers and scientists and have made it possible to extend the range, depth, and applications of MODSIM vastly, especially when the phenomena being investigated are not observable or measurements are impractical or too expensive. According to a National Science Foundation Blue-Ribbon Panel, MODSIM-based engineering and science (1) is an equal and indispensable partner, along with theory and experiment, in the quest for enhanced technological innovation; (2) holds great promise for the pervasive advancement of knowledge and understanding through discovery; (3) is indispensable to the nation's continued leadership in innovation and economic global competitiveness; and (4) is "key" to advances in a variety of fields − biomedicine, manufacturing, systems engineering, nanotechnology, health care, atmospheric and climate science, energy and environmental sciences, advanced materials, and product development. MODSIM-based engineering and science is also essential to the success of NASA's research, missions, and projects.

The MODSIM World Conference and Expo began in 2007 when the Hampton Roads Partnership's, Center for Public/Private Partnership (CP3) saw the need to share information about and interests in the vast amount of MODSIM-based research and development occurring in the Hampton Roads region of Virginia. Because of the synergy created by the efforts of Joint Forces Command; Virginia Modeling, Analysis and Simulation Center (VMASC); Eastern Virginia Medical School (EVMS), the NASA Langley Research Center (LaRC), etc., it became obvious to the CP3 membership that there was a need to establish an "interdisciplinary" forum for sharing of MODSIM knowledge and achievement. Their efforts created MODSIM World Conference & Expo. The MODSIM Conference & Expo is now in its fourth cycle.

MODSIM World 2010 was held in Hampton, Virginia, October 13−15, 2010. The theme of the 2010 conference & expo was "21st Century Decision-Making: The Art of Modeling& Simulation". The conference program consisted of seven technical tracks – Defense, Engineering and Science, Health & Medicine, Homeland Security & First Responders, The Human Dimension, K-20 STEM Education, and Serious Games & Virtual Worlds. Selected papers and presentations from MODSIM World 2010 Conference & Expo are contained in this NASA Conference Publication (CP). Section 8.0 of this CP contains papers from MODSIM World 2009 Conference & Expo that were unavailable at the time of publication of NASA/CP-2010-216205 *Selected Papers Presented at MODSIM World 2009 Conference and Expo*, March 2010.

As a condition for inclusion in the conference proceedings, the first author was responsible for securing/obtaining all permissions associated with the general release and public availability of the paper/presentation. Further, the first authors also had to grant NASA the right to include their work in the NASA CP. There are 62 papers and 41 presentations in this NASA CP. There are two appendices in this publication. Appendix A contains the names and affiliations of the conference organizers. Appendix B includes a description of the technical tracks and the names of the individuals who organized each track. Preparing the proceedings of this conference required the collaborative efforts of many individuals.

The preparation of NASA/CP-2011-217069 would not have been possible without the expertise and contributions of Ms. Leanna "Dee" Bullock (ATK Space Systems, INC.) and Ms. Jennifer McNamara (BreakAway, LTD.).

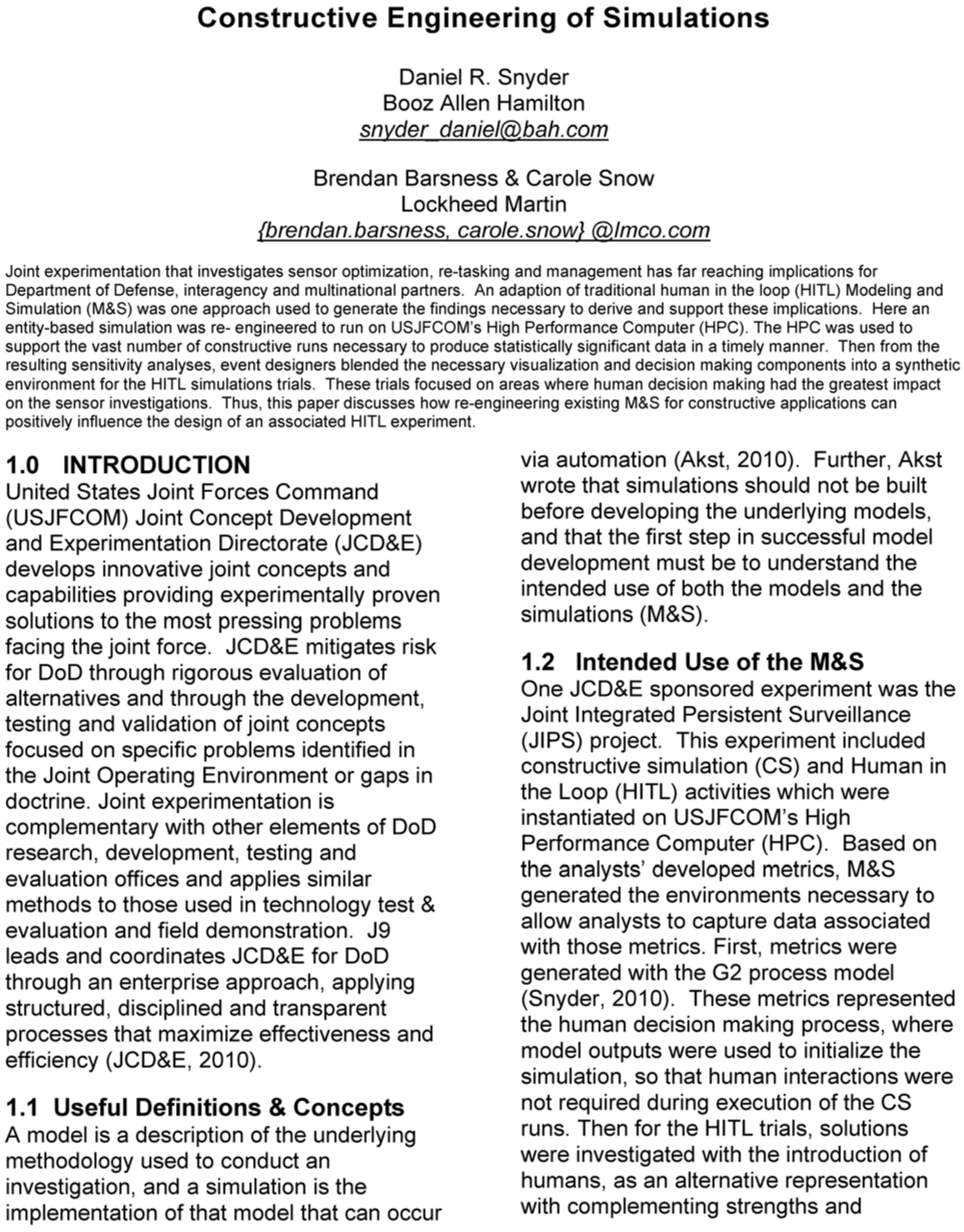
Table of Contents

Part 1

Part 2

1.0 DEFENSE TRACK

1.1 Constructive Engineering of Simulations



Constructive Engineering of Simulations

Daniel R. Snyder
Booz Allen Hamilton
snyder_daniel@bah.com

Brendan Barsness & Carole Snow
Lockheed Martin
{brendan.barsness, carole.snow} @lmco.com

Joint experimentation that investigates sensor optimization, re-tasking and management has far reaching implications for Department of Defense, interagency and multinational partners. An adaption of traditional human in the loop (HITL) Modeling and Simulation (M&S) was one approach used to generate the findings necessary to derive and support these implications. Here an entity-based simulation was re- engineered to run on USJFCOM's High Performance Computer (HPC). The HPC was used to support the vast number of constructive runs necessary to produce statistically significant data in a timely manner. Then from the resulting sensitivity analyses, event designers blended the necessary visualization and decision making components into a synthetic environment for the HITL simulations trials. These trials focused on areas where human decision making had the greatest impact on the sensor investigations. Thus, this paper discusses how re-engineering existing M&S for constructive applications can positively influence the design of an associated HITL experiment.

1.0 INTRODUCTION

United States Joint Forces Command (USJFCOM) Joint Concept Development and Experimentation Directorate (JCD&E) develops innovative joint concepts and capabilities providing experimentally proven solutions to the most pressing problems facing the joint force. JCD&E mitigates risk for DoD through rigorous evaluation of alternatives and through the development, testing and validation of joint concepts focused on specific problems identified in the Joint Operating Environment or gaps in doctrine. Joint experimentation is complementary with other elements of DoD research, development, testing and evaluation offices and applies similar methods to those used in technology test & evaluation and field demonstration. J9 leads and coordinates JCD&E for DoD through an enterprise approach, applying structured, disciplined and transparent processes that maximize effectiveness and efficiency (JCD&E, 2010).

1.1 Useful Definitions & Concepts

A model is a description of the underlying methodology used to conduct an investigation, and a simulation is the implementation of that model that can occur via automation (Akst, 2010). Further, Akst wrote that simulations should not be built before developing the underlying models, and that the first step in successful model development must be to understand the intended use of both the models and the simulations (M&S).

1.2 Intended Use of the M&S

One JCD&E sponsored experiment was the Joint Integrated Persistent Surveillance (JIPS) project. This experiment included constructive simulation (CS) and Human in the Loop (HITL) activities which were instantiated on USJFCOM's High Performance Computer (HPC). Based on the analysts' developed metrics, M&S generated the environments necessary to allow analysts to capture data associated with those metrics. First, metrics were generated with the G2 process model (Snyder, 2010). These metrics represented the human decision making process, where model outputs were used to initialize the simulation, so that human interactions were not required during execution of the CS runs. Then for the HITL trials, solutions were investigated with the introduction of humans, as an alternative representation with complementing strengths and weaknesses to the CS approach, which assisted in validating the modeled representations of the decision-making process. The CS provided a means to conduct a high number of trials not only for increased precision in the statistical output, but also to manipulate an increased number of solution set variables. Within the HITL, the number of trials was significantly decreased; however, the advantage was in the interaction of real people with the solutions. Insights that the analysts gained from the CS runs were used to assist event designers in determining the necessary M&S configurations for the associated JIPS HITL trials. Results from the HITL were used to assist designers in determining the setup of additional CS runs in keeping with a model-wargame-model (W-M-W) paradigm (Kass, 2006).

1.3 Event Design

The JIPS HITL provided experimental evidence to assist in determining the effectiveness and efficiencies of solutions against current day and futuristic baselines. Prior to the HITL runs, thousands of faster than real time CS runs were executed to examine the impact of the various solutions. Then during the HITL trials, solutions were carried forward with the introduction of a human decision-making component to further examine the solutions with the introduction of command, control, communication, computer, and intelligence (C4I) systems.

1.3.1 Experimentation Audience

Figure 1 illustrates the four echelons of command that were emulated during the HITL. These echelons were: national to include combatant command (COCOM), joint task force (JTF), division (DIV) and regimental combat team (RCT) with associated tactical elements such as battalion and companies. In turn, parts of seven levels of Command and Control (C2) were emulated by these four echelons. The blue shaded area represented the experiment audience, and the pink shaded area represented the control cell with role players. The control personnel, called the White Cell (WC), ensured that the simulation rendering of the enemy activities, friendly assets and neutral ground activities were sufficient to stimulate the participants' behavior. The analysts observed these behaviors and gathered data from the

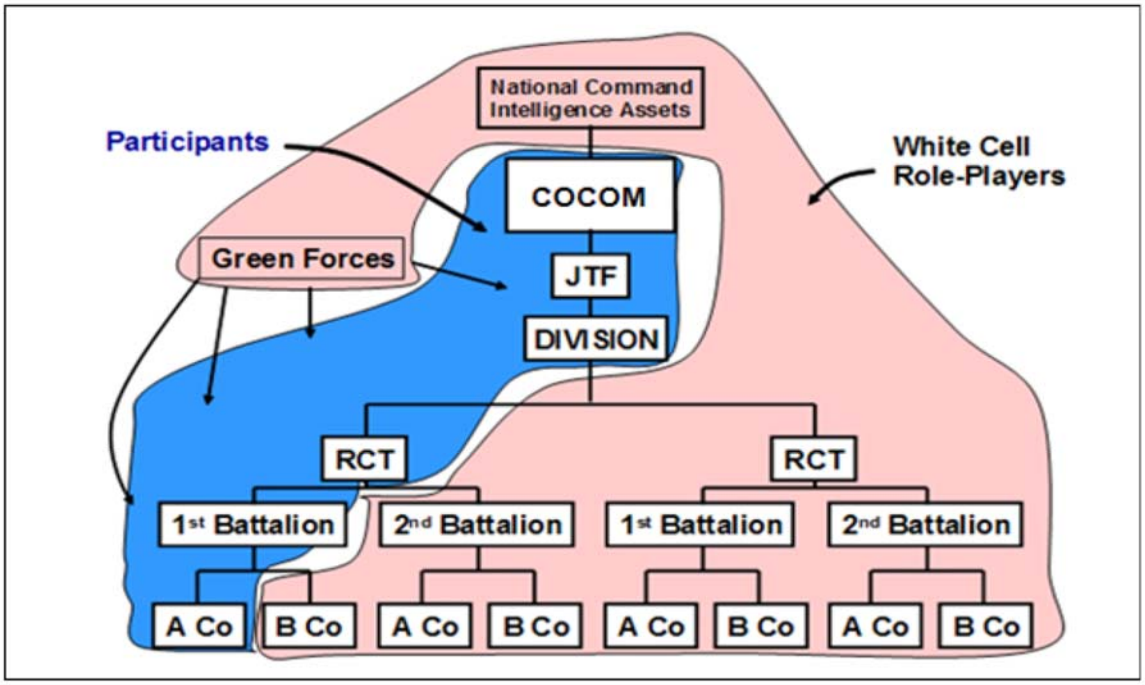


Figure 1. Separation of participants and simulation

simulation to develop their findings in keeping with the Data Collection and Analysis Plan (DCAP). M&S operators resided in the WC to assist in masking the M&S activities from the participants in order to enhance the realism of the experiment.

1.3.2 Communications

In addition to providing data in support of analysis, the simulation was also the data source for the integrated C4I systems which were used by experiment participants to carry out their duties. The M&S was used to facilitate sensor asset re-taskings, full motion video (FMV) taskings and system status reporting. Internet Relay Chat (mRIC) was the primary mode of communication among M&S operators, analysts, collection managers, targeteers and sensor asset managers at all echelons. Command and Control Personal Computer (C2PC) and Command Post of the Future (CPOF) presented participants with a shared representation of battle space entities as generated by the simulation. Email was the secondary communication channel that was available at all four echelons to send and receive lengthy textual reports such as formatted intelligence reports. Additionally, voice over IP (VOIP) provided a digital voice communication capability at all JIPS HITL workstations. These HITL workstations emulated those of real world C2 centers.

2.0 BODY

Supporting the HITL were federation M&S components that stimulated the C4I tools. These stimuli evoked behaviors from the participants which were of interest to the analysts.

2.1 Federation Components

There were six primary M&S components used in the HITL to generate situational awareness (SA) for the participants, and five of these components directly drove the Common Operating Picture (COP). First, Joint Semi-Automated Forces (JSAF) was the entity level simulation that represented the opposing entities. These entities were controlled by human operators and the JSAF automated behaviors. Second, CultureSim, a scalable entity level federate tightly coupled with JSAF that was formerly known as ClutterSim, represented the neutral populace reflected within the urban battlespace (Speicher, 2004). Third, Simulation of the Locations and Attack of Mobile Enemy Missiles (SLAMEM) was the entity level simulation used to represent sensors, sensor platforms, fusion, and tracking. Fourth, a redesigned Track Database (TrackDB) federate enabled direct communications from JSAF to the C4I components. This modified TrackDB federate replaced the need for a separate C4I gateway to feed the Global Command Control System – Joint (GCCS-J) server. C2PC and CPOF clients pulled data from GCCS-J. Fifth was the three dimensional JSAF viewer (JStealth), which produced effects that closely resembled those from emulated platforms generating FMV feeds. Finally, an experimentation & event tool suit (EETS) component, the Event Generator (EventGen), facilitated the injection of master scenario event list (MSEL) items. The EventGen injects supplemented the perceived behavior of the simulation entities. These injects were represented as emails that did not directly influence the COP, but did provide amplifying context and intent to further influence the players' reaction to the M&S stimulation of the COP.

2.2 C4I and M&S Architectures

M&S operational requirements for the JIPS HITL were broken out into the five areas. In turn, four technical spirals and one operational test were used to ensure that these five areas were sufficient to successfully accomplish the experiment from a technical standpoint (JIPS, 2010). These tests were conducted to ensure that surveillance and network architectures, simulation and communication components, plus the control procedures were sufficient to meet event objectives. Figure 2 depicts how the M&S components interfaced with the COP and the FMV view generators to stimulate the participants. Not all message traffic was processed through the TrackDB. Over-the-Horizon (OTH) Targeting Gold (OTG) messages passed initial detections directly from SLAMEM to the GCCS-J server. The perceived current track states were sent via the TrackDB. Objects and interactions from SLAMEM and the TrackDB were pulled from the Federation and placed into a MySQL repository for later analytical use. In some cases, SLAMEM data were captured in SLAMEM logs, and other Federation traffic was pulled via a separate application called JLogger. TrackDB was enhanced for this event to send track data directly to the GCCS system, removing the need for a separate C4I gateway. The JSAF Surveillance Tool was the mechanism used to allow re-tasking of sensor assets.

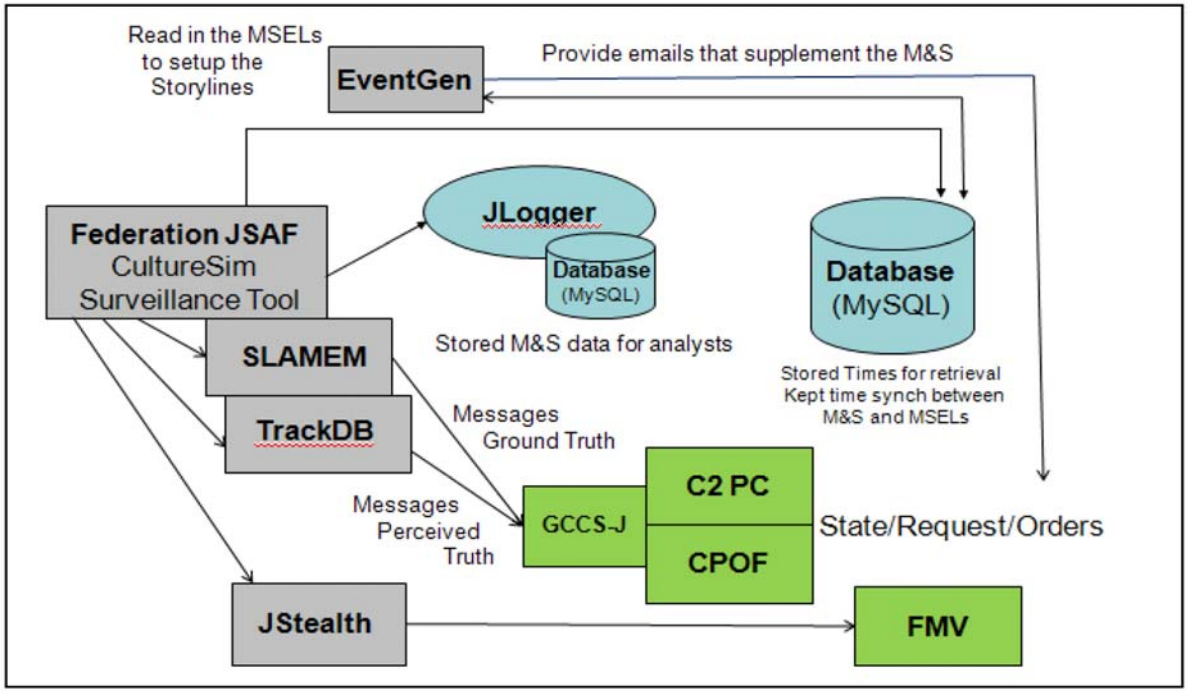


Figure 2. Simulation and C4I Architectures

2.2.1 Re-tasking Sensors Assets

During the HITL, participants tasked sensor assets to collect information on perceived enemy activities as instantiated in JSAF. Closely coupled with JSAF was the CultureSim federate which generated thousands of neutral entities which served to obscure the enemy activities that were represented in JSAF. MSEL items established the initial taskings for the sensor assets that were represented in SLAMEM. Once re-tasking requests were approved and authorized, HITL participants required the capability to influence airborne sensor assets to change the location of interest view, the assigned sensor on location, and/or fly a new route to meet a new or modified collection requirement. From an M&S perspective, blue sensor assets, represented in SLAMEM, had to be capable of changing their station, their surveillance area, or their sensor modes as directed by HITL participants.

2.2.2 Observe FMV

From an M&S perspective, FMV was simulated by attaching JStealth to simulated sensor assets that were represented in SLAMEM. The simulated FMV feeds, which were visualized through JStealth, were fed to a video server capable of streaming the video over the network to participants' C4I displays. HITL participants requested FMV feeds via the WC. Working within the WC, the M&S operators changed the FMV views per the participants' requests.

2.2.3 Inject Scenario/MSEL Events

The experiment executed a scenario with multiple storylines to provide target activity for persistent and non-persistent surveillance. Detailed events from each storyline were compiled into a single chronologically ordered event list known as the MSEL, which was executed at the discretion of the white cell. The WC controlled MSEL inject flow via EventGen. The control group also had the responsibility to add to or delete from the MSEL injects to ensure the experiment objectives were accomplished.

2.2.4 Situational Awareness

Situational awareness (SA), as applied to the JIPS HITL, was the capability to extract meaningful activities and patterns from the battlespace picture and to share this awareness across the network with appropriate participants (Hayes, 2006). The COP provided the cognitive understanding or interpretation of the JIPS battlespace. The JIPS battlespace was composed of the following four SA components (JIPS, 2010):

- Blue Force SA component – a complete picture of all Blue force entities including their identification and position. Blue force sensor entities were modeled in SLAMEM.
- Red Force SA component – a complete picture of all Red (hostile) force entities including their identification and position. Red force entities were modeled in JSAF.
- Green Force SA component – a complete picture of all civilian units/entities including their position and (optional) identification information. The Green component was modeled in JSAF, which included cultural activities, i.e., people, vehicles, and landmarks.
- Sensor SA component – a picture of the perceived entities sensed in the battlespace and reported by Blue force sensor systems. Blue force sensor reports were produced by SLAMEM.

2.2.5 M&S Data Collection

The JIPS HITL exercised the innate distributed collection mechanisms of the tools and logged all federation simulation traffic. The After Action Review (AAR) was supported by various means of consolidating the data from the HPC, data fusion techniques, and data reduction into user friendly formats as specified by the JIPS analysts.

3.0 DISCUSSION

3.1 Synthetic HITL Environment

The JIPS HITL federation was composed of simulations selected to best meet the multi-echelon command structure of joint persistent surveillance operations. The goal of the JIPS technical and operational test program was to mitigate the risk inherent in the use of disparate simulation and SA components by investigating and evaluating the functionality and interfaces of the components within the framework of the event objectives. For the JIPS experiment, persistent surveillance was defined as a collection strategy that emphasizes the ability of collection systems to linger on demand in a particular location to detect, locate, characterize, identify, track, target and assess in real or near real time (HQDA, 2010).

3.2 M&S Architecture

Various optimization techniques where employed to take advantage of the HPC architecture and address experimentation objectives.

3.2.1 Entity Count

In order to reach the JIPS HITL 100,000 entity count target, several federation routers and CultureSim nodes where configured to take advantage of the M&S Data Distribution Management (DDM) protocol. This approach reduced the number of interactions within the federation to a degree sufficient to support pockets of densely populated urban terrain areas.

3.2.2 Gateway Redesign

Given the probability that sensor collection plans would include fly-over of these dense populations, it was necessary to optimize the C4I Gateway to support the simulation-to-C4I translation of large numbers of perceived tracks. This was resolved by, 1) redesigning the TrackDB federate to replace the C4I Gateway, effectively removing the less efficient runtime database access requirement and, 2) bundling and distributing consolidated sensor reports as opposed to individual tracks.

3.2.3 Data Collection

During runtime, a distributed data collection approach was used. Collection was localized at each simulation/HPC node reducing the runtime data throughput requirement. After each HITL run, the data was off-loaded from the HPC onto a centralized repository for analysis. Colors used in figure 3 highlight the complexity of the nodal scheme: yellow blocks represented JStealth nodes, green blocks represented CultureSim nodes, purple blocks represented SLAMEM nodes, grey represented router nodes, red blocks represented JSAF nodes, and the white blocks represented the simulation that held the track histories. The node reference numbers within the blocks corresponded to 256 nodes that were mapped to a particular application. Blocks without node references represented stand alone components. Due to the heavy data traffic on the HPC, several nodes were dedicated to routing the traffic. Additionally, a majority of the HITL nodes were dedicated to instantiating JStealth and CultureSim due to their high computational demands.

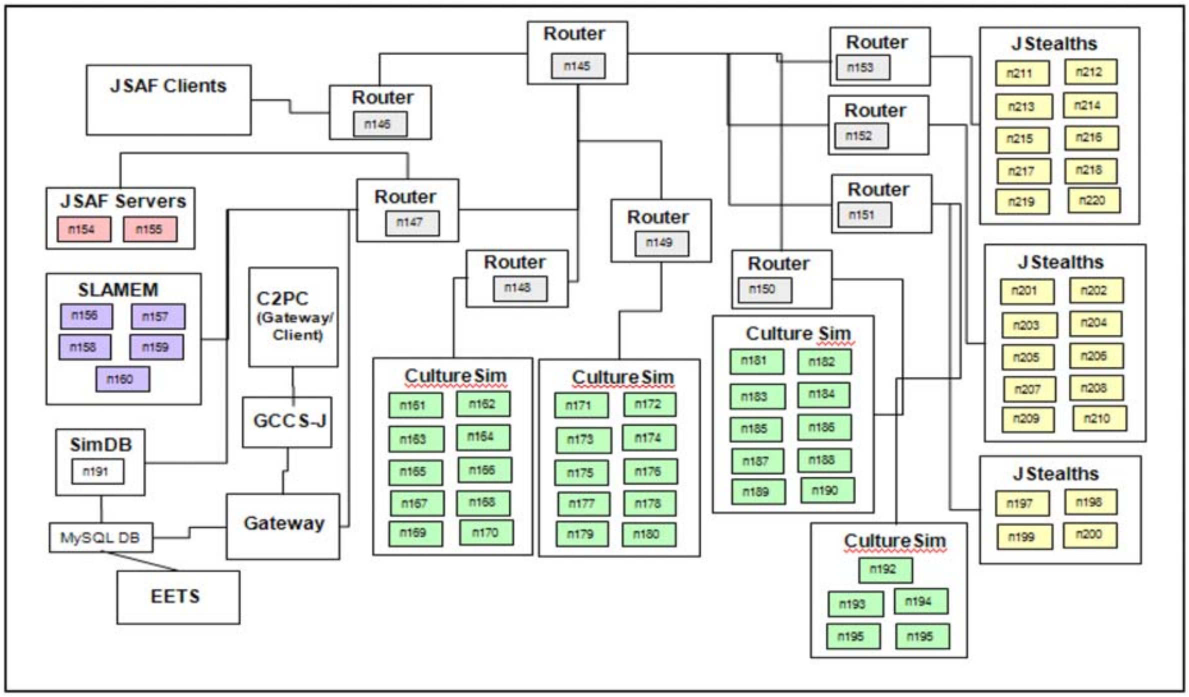


Figure 3. HPC Network Connectivity Map

3.3 Operational M&S Requirements

The approach to operational testing began with identifying a set of operational requirements having functional dependencies on the M&S. In turn, the CS runs were used to assist the setup of the HITL simulation trials. The primary difference between the CS and HITL runs was the human interactions by the participants. Observations made of these actions were used to validate the human decision making models. Other differences were associated with the addition of M&S and C4I to stimulate components that provided credible visualization and decision support aids to assist the participants in making decisions.

3.3.1 Constructive Inputs

From the CS, insights into how the FMV views should be synchronized with the MSEL injects resulted in multiple screen views for the participants. The information gathering and fusion processes, associated with the FMV views, were implicitly performed during the CS runs. By using the same sensor simulation for both the HITL and CS runs, this commonality assisted event designers in establishing traceability between the CS and HITL results.

3.3.2 Scenarios

Traceability was reinforced by using the MSEL descriptions of the scenario's storylines to develop the synthetic representations of the people, enemy forces and the necessary landmarks to stimulate the participants such that they would re-task assets as SA dictated. In other words, the JIPS event designers used the implicitly represented actions in the CS runs to determine the appropriate actions that the human participants would explicitly need to execute during the HITL runs. Simulation time was maintained in a MySQL database (Figure 2). The EventGen tool kept the MSELs synchronized with the simulations by reading and writing to this database.

3.3.3 Refining Models

Per the M-W-M paradigm, the first set of G2 models were based on expert judgment to develop time estimates that predicted the participants' performance during execution of each solution in the HITL. Next, HITL observations were used to update the experts' time estimates, which led to refining the G2 models for the second series of model runs.

4.0 CONCLUSIONS

The JIPS HITL was conducted to provide experimental evidence on the effectiveness and efficiencies of potential solutions to the persistent surveillance problem. The implementation of the HITL experiment on the USJFCOM HPC was necessary to fully implement the M-W-M paradigm. With the ability to rapidly refine the G2 models based on metrics derived from the HITL, the HPC made the re-running of the CS feasible within the project's time constraints. Thus, event designers used the HPC to achieve a closer coupling between the modeling and wargaming trials than what was previously possible in other JCD&E M&S supported activities. This coupling came from using the output data of the HITL to further refine the G2 models used for a second set of CS runs.

4.1 Constructive Engineering

Based on the results of using a version of SLAMEM in a constructive simulation mode, lessons were learned which influenced how best to custom build a JCD&E experiment with M&S and C4I components. This environment was sufficient to provide participants with the SA necessary to make meaningful decisions in line with the desires of the analysts. Observation tools and surveys were used daily, and the simulation generated metrics provided context which allowed traceability back to the DCAP. The use of common MSELs, between the CS and HITL trials, became an important factor for facilitating traceability as well as saving the time necessary to create a different set of MSELs.

4.2 Resources

Without the M&S environment created by the USJFCOM HPC, dozens of additional stand alone servers and operators would have been needed to correlate the answers and findings. Additionally, from the resulting sensitivity analyses event designers were able to select the sufficient type of visualization and decision making components to create the necessary HITL environment. It was the application of the

USJFCOM HPC resource that allowed for the focused and timely manipulation of simulation data, routing of data traffic, and stimulation of the C4I systems that provided the human decision makers a credible environment for JIPS investigations. This manipulation of data was sufficient to provoke the participants to take actions that were of interest to the analysts.

4.3 Optimization Techniques

Intelligently exploiting the M&S DDM protocol on the HPC led to the reduction of inter-node simulation interactions thus reducing network traffic and unnecessary message processing. Additionally, localizing the collection of data at each simulation/HPC node reduced the runtime data throughput requirement. After each HITL run, the data was off-loaded from the HPC onto a centralized repository for analysis. However, it was the redesign of the TrackDB federate and bundling of sensor reports that proved to provide the most benefit to the participants. This was because the redesign and reconfiguration reduced C4I track latencies, and contributed most to an environment that dispelled disbelief in a constructively engineered world.

5.0 REFERENCES

[1]. Akst, G. (2010) M&S Is Not One Word! , Alexandria, Virginia: *Military Operations Research Society, Phalanx, Volume 43, No. 2*, from www.mors.org.

[2]. Alberts, D. & Hayes, R. (2006) Understanding Command and Control, Washington, DC: *Command and Control Research Program (CCRP),* from www.dodccrp.org.

[3]. Guiding Principles for Joint Concept Development and Experimentation (JCD&E) (2010). USJFCOM J9, from www.jfcom.mil.

[4]. Headquarters, Department of the Army (HQDA), (2010). *Field Manual 2-0 Intelligence,* from www.us.army.mil.

[5]. Kass, R. (2006). The Logic of Warfighting Experiments, Washington, DC: *Command and Control Research Program (CCRP),* from www.dodccrp.org.

[6]. Joint Integrated Persistent Surveillance (JIPS) Modeling & Simulation (M&S) Operational Test Strategy (2010). USJFCOM J9, from www.jfcom.mil.

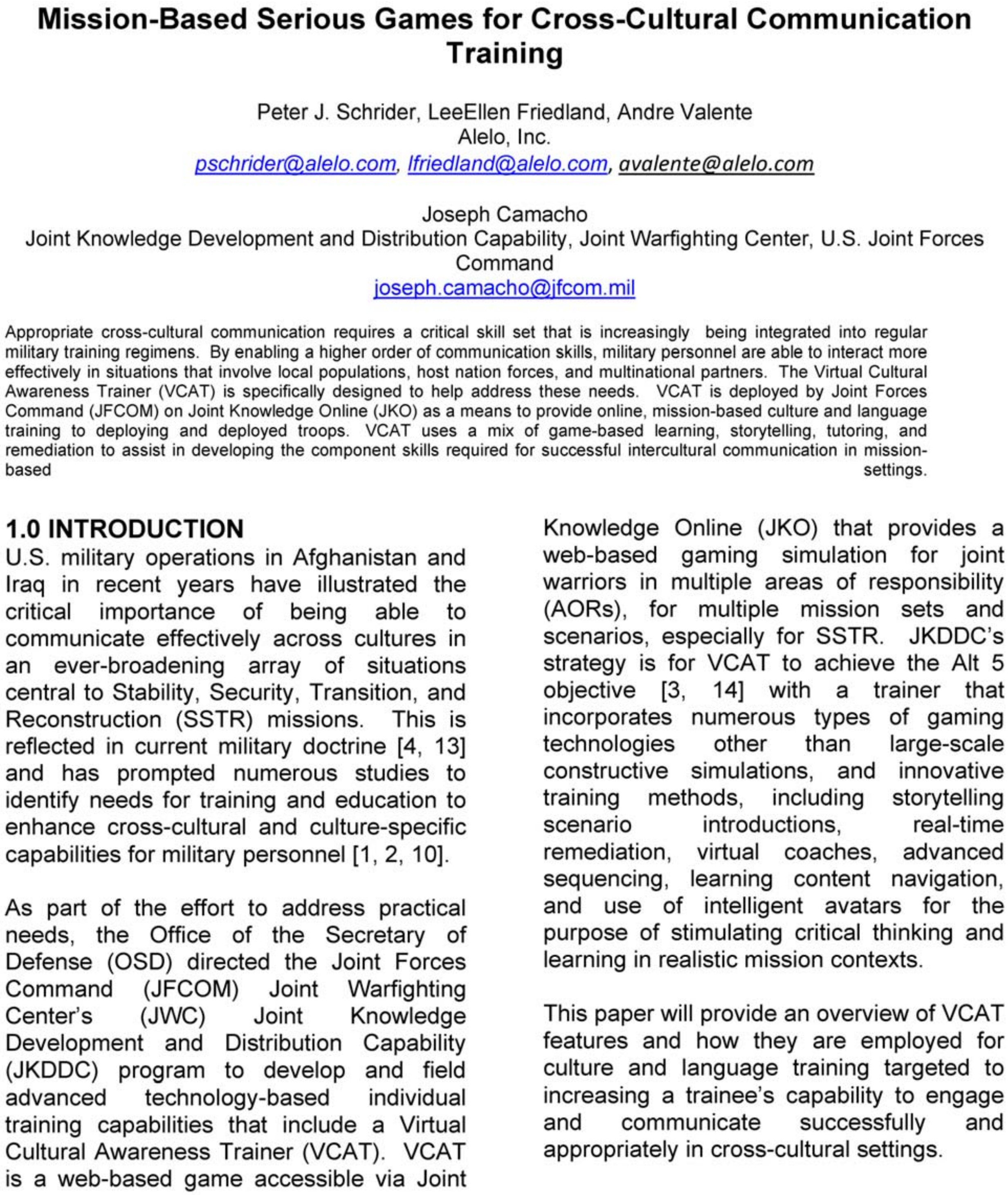
[7]. Snyder D. & Brewton, C. (2010). Pointing the Way with Constructive Simulations; in press. *Proceedings of the 2010 Interservice/Industry Training Simulation and Education Conference.*

[8]. Speicher D. & Wilbert D. (2004). Simulating Urban Traffic in Support of the Joint Urban Operations Experiment. *Proceedings of the 2004 Interservice/Industry Training Simulation and Education Conference*.

6.0 ACKNOWLEDGMENTS

This material is based an experiment sponsored by the USJFCOM, and supported by a contract with the Lockheed Martin Corporation, Toyon, SimIS Inc., KES Inc., Information Sciences Institute, and Booz Allen Hamilton. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of USJFCOM or the U.S. Government.

1.2 Mission-Based Serious Games for Cross-Cultural Communication Training



Mission-Based Serious Games for Cross-Cultural Communication Training

Peter J. Schrider, LeeEllen Friedland, Andre Valente
Alelo, Inc.
pschrider@alelo.com, lfriedland@alelo.com, avalente@alelo.com

Joseph Camacho
Joint Knowledge Development and Distribution Capability, Joint Warfighting Center, U.S. Joint Forces Command
joseph.camacho@jfcom.mil

Appropriate cross-cultural communication requires a critical skill set that is increasingly being integrated into regular military training regimens. By enabling a higher order of communication skills, military personnel are able to interact more effectively in situations that involve local populations, host nation forces, and multinational partners. The Virtual Cultural Awareness Trainer (VCAT) is specifically designed to help address these needs. VCAT is deployed by Joint Forces Command (JFCOM) on Joint Knowledge Online (JKO) as a means to provide online, mission-based culture and language training to deploying and deployed troops. VCAT uses a mix of game-based learning, storytelling, tutoring, and remediation to assist in developing the component skills required for successful intercultural communication in mission-based settings.

1.0 INTRODUCTION

U.S. military operations in Afghanistan and Iraq in recent years have illustrated the critical importance of being able to communicate effectively across cultures in an ever-broadening array of situations central to Stability, Security, Transition, and Reconstruction (SSTR) missions. This is reflected in current military doctrine [4, 13] and has prompted numerous studies to identify needs for training and education to enhance cross-cultural and culture-specific capabilities for military personnel [1, 2, 10].

As part of the effort to address practical needs, the Office of the Secretary of Defense (OSD) directed the Joint Forces Command (JFCOM) Joint Warfighting Center's (JWC) Joint Knowledge Development and Distribution Capability (JKDDC) program to develop and field advanced technology-based individual training capabilities that include a Virtual Cultural Awareness Trainer (VCAT). VCAT is a web-based game accessible via Joint Knowledge Online (JKO) that provides a web-based gaming simulation for joint warriors in multiple areas of responsibility (AORs), for multiple mission sets and scenarios, especially for SSTR. JKDDC's strategy is for VCAT to achieve the Alt 5 objective [3, 14] with a trainer that incorporates numerous types of gaming technologies other than large-scale constructive simulations, and innovative training methods, including storytelling scenario introductions, real-time remediation, virtual coaches, advanced sequencing, learning content navigation, and use of intelligent avatars for the purpose of stimulating critical thinking and learning in realistic mission contexts.

This paper will provide an overview of VCAT features and how they are employed for culture and language training targeted to increasing a trainee's capability to engage and communicate successfully and appropriately in cross-cultural settings.

2.0 VCAT AND SERIOUS GAMES

In order to fully understand the value of mission-based serious games, we must first understand what a serious game is and the impact and value of game-based training. A serious game is “the digital application of gaming technology, process and design to the solution of problems faced by business, government, academia and other organizations” [6]. Serious games use the same technology that is used in the entertainment video game industry, but are designed to focus on enabling the player, or trainee, to achieve a specific business- or goal-driven training outcome. Successful serious games provide trainees with immersive and engaging training [5, 12] that can be delivered on a variety of platforms, including the PC, across the Internet, and on personal handheld devices. Serious games provide a unique combination of challenge and engagement in a single training tool.

Research has shown that trainees are impacted positively by the time they spend playing a training game, as well as by their level of motivation and satisfaction when using game-based training. A more motivated and satisfied trainee will become more highly engaged. They allow more time and attention to the training and thereby become more skilled as a result [11].

In general, there are two types of serious games for training: broad-based mission operational training and task-specific training. Both types of games provide benefits to the training audience.

Task-based training is based on an analysis of the performance components required to complete a given task and the skills needed to execute task elements [9]. For example, a trainee may be required to know how to replace a specific component on an engine or how to take appropriate actions to control a crowd. The effects of the situation or context in which a task is to be performed may also have implications for training, as fixing an engine in a shop versus a desert may require the mechanic to adapt and improvise, and controlling a friendly crowd of children versus an angry crowd of protesters may require the security officer to employ different techniques.

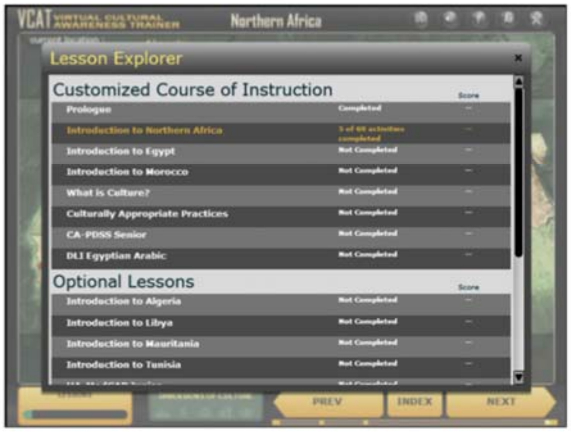
In contrast, mission-based training is directed more toward achieving a complex developmental goal that will usually involve implementing several tasks simultaneously and may not have a specific predefined outcome. Mission-based training tends to focus on more attitudes, knowledge, and skills collectively rather than just the skills that are the typical focus of task-based training.

The VCAT course combines the specific benefits of task-oriented training with the complexity of mission-based objectives to provide a comprehensive course that allows users to develop well-rounded knowledge of their situational environment, as well as their mission objectives. VCAT integrates serious games to deliver high-impact training to the user. The serious games are combined with quizzes, videos, and mini-games that allow trainees to practice and be tested on what they have learned [7].

2.1 VCAT Course Components

The VCAT curriculum is designed to provide trainees with the knowledge, skills, and attitudes they will need for intercultural interactions in order to successfully conduct missions during their deployment. The course is built with significant input from subject matter experts (SMEs), including active-duty and former military personnel, domain and occupational experts, natives of the target geographical areas, native speakers of the target languages, sociocultural and linguistic anthropologists, and instructional design experts [7, 8]. This foundation helps to ensure that training goals, as well as course information and exercises, address realistic needs and scenarios that trainees may experience.

Every VCAT course begins with a Prologue that introduces trainees to the course and the region or subject focus (such as cross-cultural competence) in question. Trainees then answer a Placement Questionnaire that captures information about a trainee's background and level. This information is used to construct a customized curriculum for every learner. Below is the lesson navigator showing the customized course of instruction plus the optional materials.



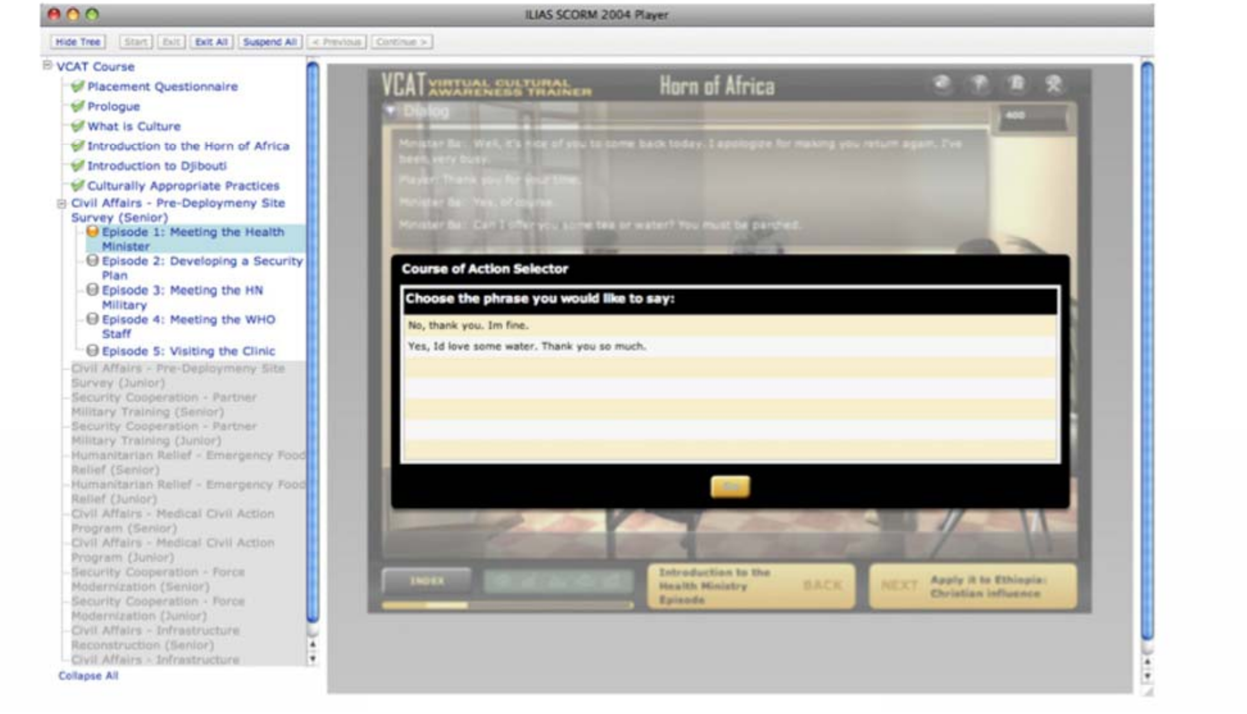
Lesson Explorer

Customized Course of Instruction		Score
Prologue	Completed	–
Introduction to Northern Africa	5 of 68 activities completed	–
Introduction to Egypt	Not Completed	–
Introduction to Morocco	Not Completed	–
What is Culture?	Not Completed	–
Culturally Appropriate Practices	Not Completed	–
CA-PDSS Senior	Not Completed	–
DLI Egyptian Arabic	Not Completed	–

Optional Lessons		Score
Introduction to Algeria	Not Completed	–
Introduction to Libya	Not Completed	–
Introduction to Mauritania	Not Completed	–
Introduction to Tunisia	Not Completed	–

The tailoring is not limited to the definition of a course of instruction. VCAT activities use a flexible multimedia instructional strategy in which users control their experience. For example, auditory learners can choose to have the coach provide narration, while visual learners can turn it off to suit their learning style.

The core VCAT curriculum starts by having trainees learn introductory knowledge about the region in question to broaden their understanding of the social and cultural factors that are most important in the focus region. Upon completion of the regional introduction, trainees enter the core episode clusters that have been selected for them based on the Placement Questionnaire. These episode clusters are structured to reflect the trainee's area of deployment, mission assignment, and role in mission activity. VCAT provides lessons with instruction for acquiring specific operations-oriented communication skills, and then provides a combination of practice tests, formal tests, and simulated intercultural encounters in which trainees can use their newly enhanced judgment and skills to perform tasks and interact appropriately in intercultural mission-specific situations [8]. The course is built to conform to the latest SCORM standards and works on the JKO's



VCAT Course
- Placement Questionnaire
- Prologue
- What is Culture
- Introduction to the Horn of Africa
- Introduction to Djibouti
- Culturally Appropriate Practices
- Civil Affairs - Pre-Deploymeny Site Survey (Senior)
  - Episode 1: Meeting the Health Minister
  - Episode 2: Developing a Security Plan
  - Episode 3: Meeting the HN Military
  - Episode 4: Meeting the WHO Staff
  - Episode 5: Visiting the Clinic
- Civil Affairs - Pre-Deploymeny Site Survey (Junior)
- Security Cooperation - Partner Military Training (Senior)
- Security Cooperation - Partner Military Training (Junior)
- Humanitarian Relief - Emergency Food Relief (Senior)
- Humanitarian Relief - Emergency Food Relief (Junior)
- Civil Affairs - Medical Civil Action Program (Senior)
- Civil Affairs - Medical Civil Action Program (Junior)
- Security Cooperation - Force Modernization (Senior)
- Security Cooperation - Force Modernization (Junior)
- Civil Affairs - Infrastructure Reconstruction (Senior)
- Civil Affairs - Infrastructure

Minister Ba: Well, it's nice of you to come back today. I apologize for making you return again. I've been very busy.

Player: Thank you for your time.

Minister Ba: Yes, of course.

Minister Ba: Can I offer you some tea or water? You must be parched.

Course of Action Selector

Choose the phrase you would like to say:

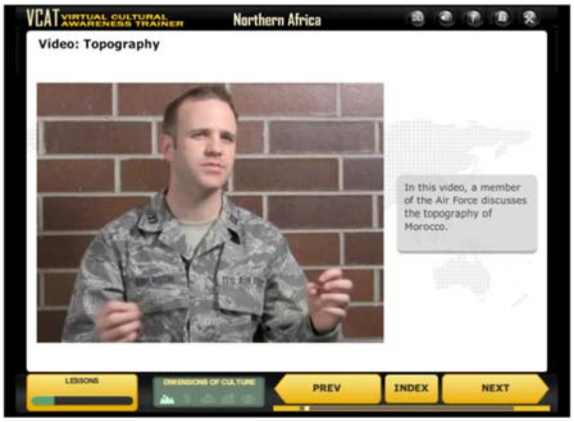
No, thank you. Im fine.

Yes, Id love some water. Thank you so much.

Introduction to the Health Ministry Episode

Apply it to Ethiopia: Christian influence native AtlasPro Learning Management System (LMS).

VCAT incorporates a variety of media to help keep the trainee engaged and interested in the course materials. For example, videos and game-based scenarios are embedded into lessons where the user is presented with a dynamic culture-based situation. In this simulation, users are forced to draw upon their knowledge of local customs and behavior, as well as their ability to interact with intelligent human characters in culturally-appropriate ways.

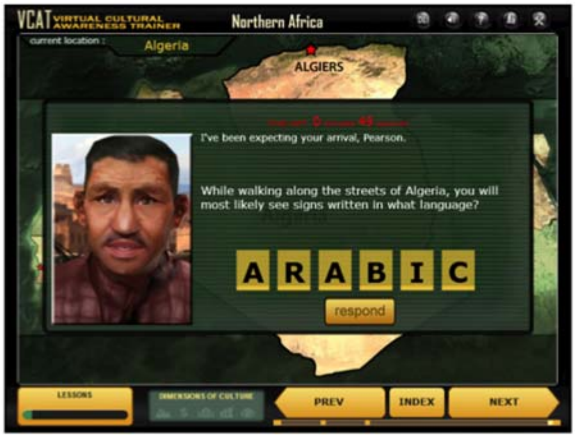


Video: Topography

In this video, a member of the Air Force discusses the topography of Morocco.

One example of a serious game in VCAT is a mini-game, called Culture Quest. In this game, trainees have to answer questions correctly and, when they do, they are rewarded with a tile. As the game continues, the player accumulates multiple tiles that eventually can be combined into a picture puzzle. Once the picture puzzle is fully assembled, the player is surprised by a video photo montage that flashes pictures portraying cultural scenes from across the region of study. This game provides a mechanism to reinforce the new knowledge trainees have learned in their lessons and provide a new and fun way to use what they've learned.

Another serious game element in VCAT is seen in interactive scenarios that are embedded throughout the course. These scenarios present trainees with a game-based video scene in which the player has to select a course of action. Trainees pick an action and their character in the game then role-plays the behavior associated with that choice. The non-player characters in the scenario then react to the player behavior. Following that, an interactional sequence ensues that requires the player to make additional appropriate choices in order to resolve the interaction satisfactorily according to local cultural norms and mission goals. This game-based activity enables trainees to practice applying the knowledge and skills they have acquired through the lessons in a real-time environment reflecting cultural, task, and mission factors. Key in this activity is the need for trainees to make decisions about how to proceed in response to different types of unfamiliar, desirable, or undesirable reactions from non-player characters.

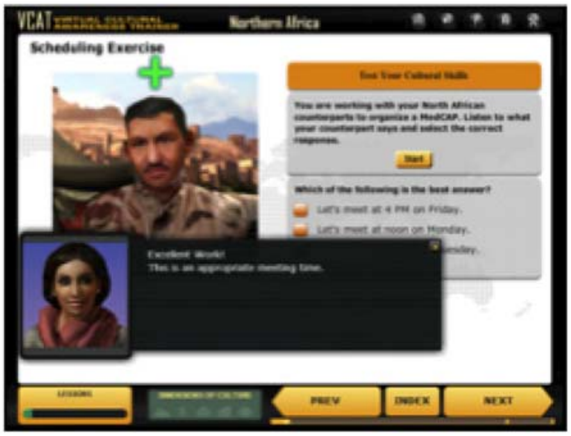


I've been expecting your arrival, Pearson.

While walking along the streets of Algeria, you will most likely see signs written in what language?

The interactive scenarios embedded throughout the course are specifically designed to address the requirements of the real-world missions and AORs that trainees must negotiate. Cross-cultural factors are always at play, including attitudes that the non-player characters have toward the trainee's character. The choices available to a trainee in each scenario vary according to how well face-to-face interaction between characters proceeds and how that progress influences the non-player characters. In other words, the trainee is immersed in a responsive and adaptive environment rather than a fixed state scene, and must negotiate through potentially ambiguous or nuanced engagements.

Interactive scenarios are available in three forms. First, mini-conversation exercises are one-move dialogs in which the learner is prompted with a verbal and gestural input from a character in the AOR and must select the appropriate response. The system provides scaffolding in the form of a "+" or "-" sign and a corresponding earcon indicating whether the character reacted positively or negatively to the learner's choice. Further, an interactive coach appears and provides both positive reinforcement as well as detailed feedback about the choice. A screenshot of a mini-conversation exercise showing positive feedback and the coach is shown below.



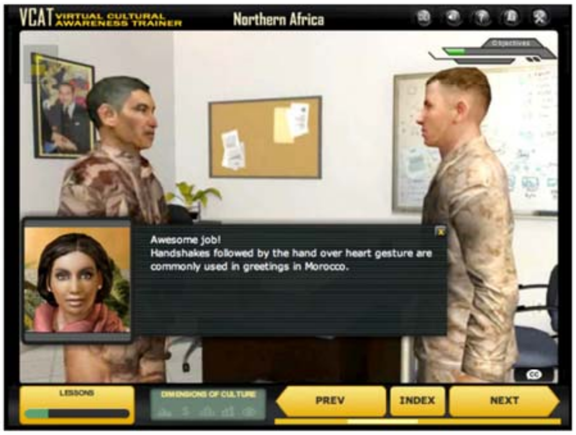
You are working with your North African counterparts to organize a MedCAP. Listen to what your counterpart says and select the correct response.

Which of the following is the best answer?

- Let's meet at 4 PM on Friday.
- Let's meet at noon on Monday.
- [illegible]

Excellent Work!
This is an appropriate meeting time.

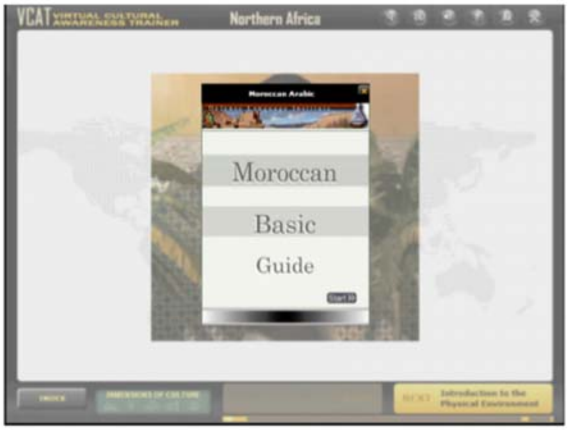
A second type of interactive scenario is a practice episode. A practice episode immerses the learner in a mission setting, provides a number of goals, and exposes the learner to a number of situations as the scenario unfolds. At each move, the learner gets the same type of scaffolded feedback described above, as well as the opportunity to request detailed feedback from the coach. At the end of a test scenario, VCAT uses the coach to provides the learner with a debriefing as well as a detailed after action review. Depending on the results of the after action review, VCAT also provides the learner with tailored remediation in the form of a custom set of lesson materials that address any subjects or skills the learner has not mastered in the scenario. The learner can and should perform practice episodes several times, exploring the consequences of specific cultural choices. Below is a screenshot of a practice episode.



Awesome job!
Handshakes followed by the hand over heart gesture are commonly used in greetings in Morocco.

Mini-conversation exercises and practice episodes provide detailed training. A third type of interactive simulation called a test episode, tests the learner's ability to put these skills into practice. Test episodes are similar to practice episodes, but the learner does not get any feedback or coaching except for an indication of which mission objectives have been completed. The test episode is used for assessing the learning outcomes. A learner needs to accomplish at least 80% of the mission objectives in order to pass a test episode and that module of VCAT.

VCAT also provides language instruction by integrating Language Survival Guides from the Defense Language Institute. The system selects the Guide for the specific language relevant to the country or region selected by the learner. For example, below is a screen shot of the lesson with the survival guide for Moroccan Arabic, offered to a learner who selected Morocco as the country of deployment.



3.0 Conclusion

The Virtual Cultural Awareness Trainer breaks new ground in a number of ways. It combines the engaging and fun elements of serious games with the power of innovative instructional design and advanced technologies that result in customizable courses of instruction. It integrates varied and interesting multimedia and game-based components into lessons and provides culture- and mission-based knowledge and perspectives based on the first-hand experience of subject matter experts, many of whom have played the same real-life roles that trainees will play when deployed. VCAT thereby provides a flexible, adaptive, and fun training resource that brings serious games into the schoolhouse.

4.0 References

1. Abbe, A. (2008). Building Cultural Capability for Full-Spectrum Operations. (ARI Study Report 2008-04). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

2. Abbe, A., Gulick, L. M. V., & Herman, J. L. (2007). Cross-Cultural Competence in Army Leaders: A Conceptual and Empirical Foundation. (ARI Study Report 2008-01). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

3. Gardner, D., Hartman, F., "Transforming Joint Training, The Office of the Deputy Undersecretary of Defense for Readiness discusses the Training Capabilities Analysis of Alternatives Study Available at http://www.t2net.org/downloads/briefs/news/MT2Issue9.6.pdf.

4. Counterinsurgency. (2006). Field Manual 3-24. Washington, DC: Headquarters, Department of the Army.

5. Gee, J. P. (2003). What Video Games Have to Teach us about Learning and Literacy. Computers in Entertainment (CIE), 1(1), 20-20.

6. I/ITSEC Serious Games Showcase and Challenge. Available at http://www.sgschallenge.com/forms/Challenge-2010-rules.pdf.

7. Johnson, W.L. (2010). Using Immersive Simulations to Develop Intercultural Competence. In Culture and Computing. Berlin: Springer-Verlag.

8. Johnson, W.L., Friedland, L., Watson, A., & Surface, E. (in press). The Art and Science of Developing Intercultural Competence. In Paula J. Durlach & Alan M. Lesgold (Eds.), Adaptive Technologies for Training and Education. New York: Cambridge University Press.

9. Jonassen, D.H., Tessmer, M., & Hannum, W.H. (1999). Task Analysis Methods for Instructional Design. Mahwah, NJ: Lawrence Erlbaum.

10. McDonald, D.P., McGuire, G., Johnson, J., Selmeski, B., & Abbe, A. (2008). Developing and Managing Cross-Cultural Competence within the Department of Defense: Recommendations for Learning and

Assessment. Technical Report, DoD RACCA Working Group.

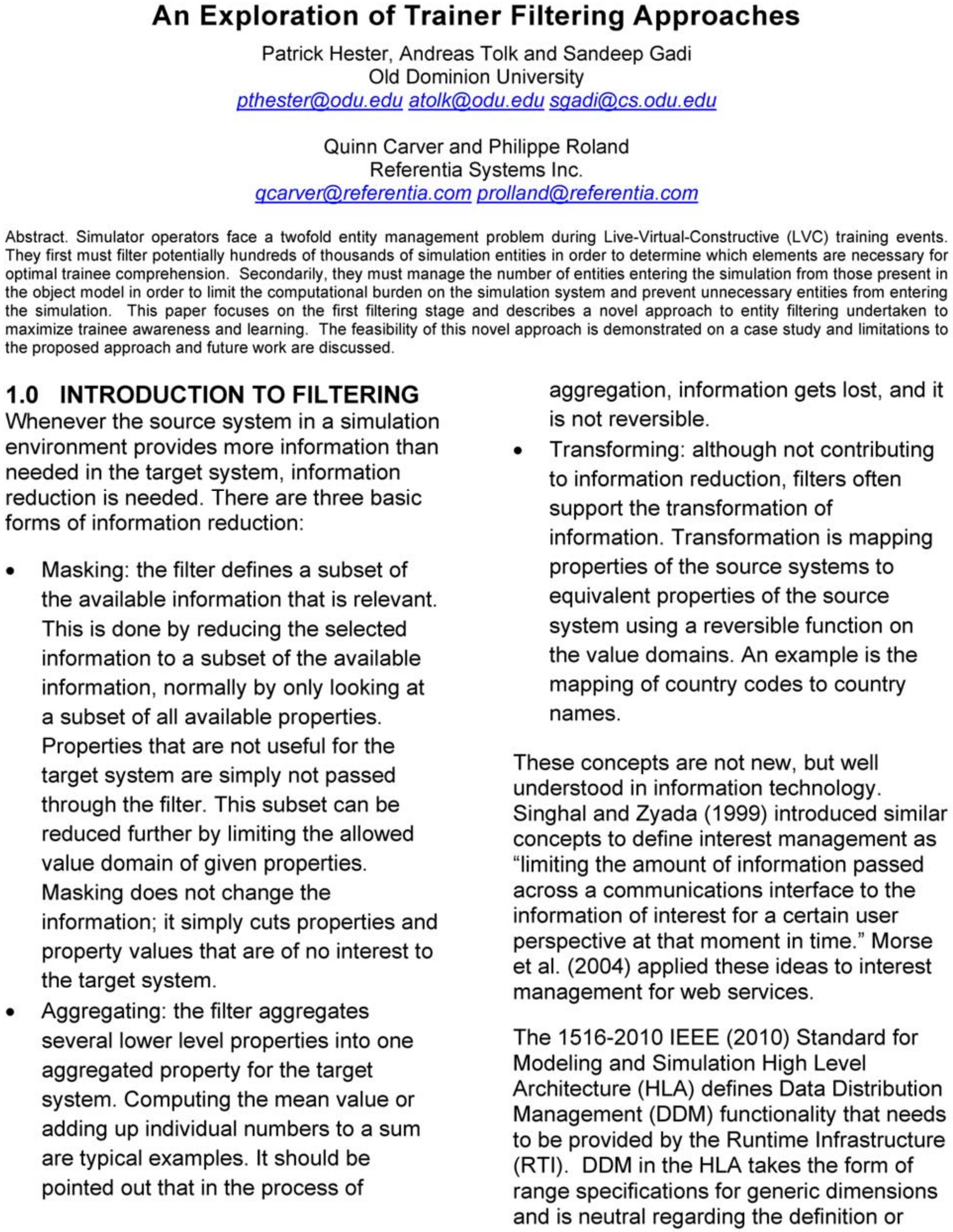
11. Orvis, Karin A., Daniel B. Horn, and James Beelanich. (2007). (ARI Technical Report 1202). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

12. Prensky, M. (2001). Digital Game-Based Learning. Columbus, OH: McGraw-Hill.

13. Stability and Support Operations. (2003). Field Manual 3-07. Washington, DC: Headquarters, Department of the Army.

14. Office of Undersecretary of Defense, Personnel and Readiness (2007). An Innovative Approach for Training Acquisitions. Available at http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA493858

1.3 An Exploration of Trainer Filtering Approaches



An Exploration of Trainer Filtering Approaches

Patrick Hester, Andreas Tolk and Sandeep Gadi
Old Dominion University
pthester@odu.edu atolk@odu.edu sgadi@cs.odu.edu

Quinn Carver and Philippe Roland
Referentia Systems Inc.
qcarver@referentia.com prolland@referentia.com

Abstract. Simulator operators face a twofold entity management problem during Live-Virtual-Constructive (LVC) training events. They first must filter potentially hundreds of thousands of simulation entities in order to determine which elements are necessary for optimal trainee comprehension. Secondarily, they must manage the number of entities entering the simulation from those present in the object model in order to limit the computational burden on the simulation system and prevent unnecessary entities from entering the simulation. This paper focuses on the first filtering stage and describes a novel approach to entity filtering undertaken to maximize trainee awareness and learning. The feasibility of this novel approach is demonstrated on a case study and limitations to the proposed approach and future work are discussed.

1.0 INTRODUCTION TO FILTERING

Whenever the source system in a simulation environment provides more information than needed in the target system, information reduction is needed. There are three basic forms of information reduction:

- Masking: the filter defines a subset of the available information that is relevant. This is done by reducing the selected information to a subset of the available information, normally by only looking at a subset of all available properties. Properties that are not useful for the target system are simply not passed through the filter. This subset can be reduced further by limiting the allowed value domain of given properties. Masking does not change the information; it simply cuts properties and property values that are of no interest to the target system.
- Aggregating: the filter aggregates several lower level properties into one aggregated property for the target system. Computing the mean value or adding up individual numbers to a sum are typical examples. It should be pointed out that in the process of aggregation, information gets lost, and it is not reversible.
- Transforming: although not contributing to information reduction, filters often support the transformation of information. Transformation is mapping properties of the source systems to equivalent properties of the source system using a reversible function on the value domains. An example is the mapping of country codes to country names.

These concepts are not new, but well understood in information technology. Singhal and Zyada (1999) introduced similar concepts to define interest management as "limiting the amount of information passed across a communications interface to the information of interest for a certain user perspective at that moment in time." Morse et al. (2004) applied these ideas to interest management for web services.

The 1516-2010 IEEE (2010) Standard for Modeling and Simulation High Level Architecture (HLA) defines Data Distribution Management (DDM) functionality that needs to be provided by the Runtime Infrastructure (RTI). DDM in the HLA takes the form of range specifications for generic dimensions and is neutral regarding the definition or meaning of the dimensions. This approach implements the masking of information. Aggregation and transformation are within the responsibility of the participating federate. DDM, however, allows individual federates to specify individual filters through which they can receive their information. It should be pointed out that the DDM specifications differ significantly between the various versions of HLA (1.3NG, IEEE 1516, IEEE 1516 Evolved) and also between different implementations of these standards. As such, DDM is standardized, but the community has not embraced a particular solution so far.

This paper explores filtering and DDM in the context of a particular problem, discussed in the following section, along with solution requirements. Following is a discussed on the developed solution. Finally, some conclusions and recommendations for future work are provided.

2.0 PROBLEM DISCUSSION

In the particular DDM problem addressed in this study the challenge is to select from thousands of constructive simulated entities and dozens of live entities that interact in a training domain those being relevant to a trainee within a virtual trainer whose graphical capabilities are limited to display only a subset, typically less than 100 entities. On the one hand side, the reason for this need for information reduction can be the limitation of the simulator due to the technical nature of the display. On the other hand side, the reason can be the need for special training, like reducing overload for new recruits by displaying too many options or the focus on a particular set of targets. In both cases, the selection of the targets to be displayed should be configurable by the trainer based on his constraints.

Among the constraints for the architecture considerations are the following:

- All pieces of information needed to display the chosen entities – or an aggregation thereof – need to be extracted from the training domain (Fleet Synthetic Training (FST) Simulation Domain).
- Only those pieces of information needed to display the chosen entities – or an aggregation thereof – should be extracted from the training domain (FST Simulation Domain).
- Aggregation of information and masking should be done at the earliest possible point, which is when it can be assured that nobody else needs the original pieces of information any longer.
- The FST Simulation Domain should not need to be modified for the solution.

The Joint Live Virtual Constructive (LVC) Data Translator (JLVCDT), also known to as JBUS, is a potential solution to provide these masking and aggregating services. JBUS has originally been developed by the USJFCOM to enable the easy integration and data translation between joint simulation system. JBUS was developed to serve as federation bridging utility, providing federation to federation connectivity, federation to external protocol connectivity, and data filtering between protocols and federations. The US Navy developed NCTE and NASMP federate object model support for JBUS to support FST. This JBUS version currently JBUS filtering utilities contains several built-in filters, reducing the burden of filter development. Filter options within JBUS are currently configured via check boxes within the JBUS GUI for filters.

An initial approach to provide more dynamic filtering envisioned a solution where JBUS filter options could be configured remotely. Filter command messages interpreted by JBUS would be sent as payload in an already existing command message. The values for the infrastructure of choice can be changed by XML messages that are used to update the respective selections corresponding to commands available in the GUI.

The definition of an XML schema for command and control of JBUS was one of the first design activities.

A Training Aware Common Operational Picture (TACOP) will provide tools for monitoring real-time data in a large scale LVC training environment. The TACOP system will be built as a filtering framework, filter controls, and an IOS interaction interface. The following section discusses the authors' solution for developing a filtering approach.

3.0 DISCUSSION

3.1 Proposed Solution

Given the constraints proposed for TACOP, the architecture shown in Figure 1 was developed by the authors.

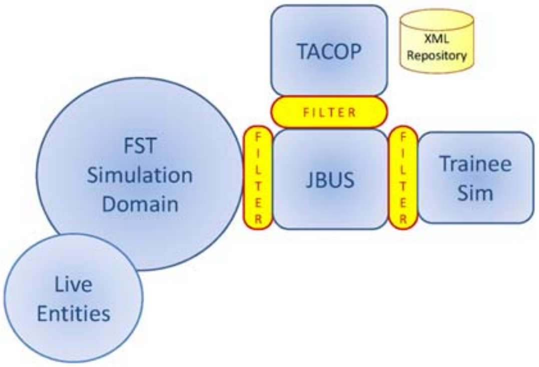


Figure 1: Initial Proposed System Architecture

The assumptions for this recommended point of view can be summarized as follows:

- The JBUS object model (JOM) itself can be interpreted as the master filter, as only elements captured in the JOM can be communicated using JBUS. This is a technical assumption.
- The FST Simulation Domain is outside of the control of the trainer. It simply provides the operational context and all relevant data in support of various training objectives. All activities the trainer is interested in for his task are embedded into a broader operational context within this domain. This is an operational assumption.
- The trainer only needs a subset of all of the information that is available in the FST Simulation Domain. The information required is training objective specific, e.g., the same scenario can be used to support air operations against submarines as well as against convoy operations. The subset of the overall scenario the trainer needs to see is different for both cases. This is an operational assumption.
- The trainer needs to see all information displayed for the trainee as well as additional information he needs to choose the next configuration to be selected for the trainee. Therefore, the trainee picture must be a real subset of the trainer picture. This is an operational assumption.

This motivated recommendations for the design of three filters:

- Filter One (between FST and JBUS) is configured to support the training objective. It filters out all elements that were not relevant for the current training objective.
- Filter Two (between JBUS and TACOP) is configured to optimize the information presentation for the trainer. Based the on current situation, this filter selects a subset of everything that passed filter one and that shall be displayed for the trainer. This filter can be used for various purposes, such as filtering out detailed information needed for 3D high resolution displays for the trainee, but that are not needed for the TACOP display (technical optimization); or to avoid cognitive overload the number of

displayed units can be limited to a maximal number close to the training area (cognitive optimization).

- Filter Three (between JBUS and Trainee) is configured to display the subset of information the trainee needs to see for optimal training purposes. It makes sense to assume that this is a subset of what the trainer sees (although this is not necessary).

All three filters are needed, and all three filters need to be configured in support of technical optimization for engineers, operational optimization for military users, and cognitive optimization for psychologists supporting the training.

3.2 Implemented Solution

The TACOP system will be built as a filtering framework, filter controls, and an IOS interaction interface. The filtering framework will accept and respond to commands and execute the actual filtering. The filter controls will present the capability and flexibility of the framework as an intuitive GUI that makes training a simpler, more manageable task for the instructor operator. TACOP will also develop an extensible protocol for communicating filtering commands and responses of the IOS to the filtering framework through the HLA network.

The goal was to then be able to build a generalized schema to accommodate the addition of new filters in the future. To build the schema, we used Liquid XML Studio 2010. Filters were categorized into filter types and each of these filters contains a name, id, parameters and description as tags. After generalizing the schema, the next step was to show the interactions between the filtering framework and IOS. For this we developed use cases to start with and chose JAXB (Oracle, 2003) to work with the xml communication between the filtering framework (backend) and IOS (frontend).

Our team decided to use JAXB, since it's a straightforward software tool in java which allows java developers to access and process XML. After generating the required classes in JAXB, we developed a frontend (IOS) and backend (filtering framework) and we now try to establish communication between these two as a prototype to the communication between the Filtering network and IOS. This simplified architecture in shown in Figure 2 below, showing also how the simplified version has been derived from the original concepts

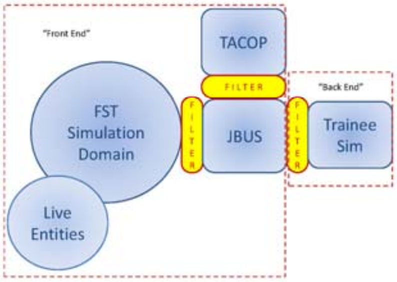
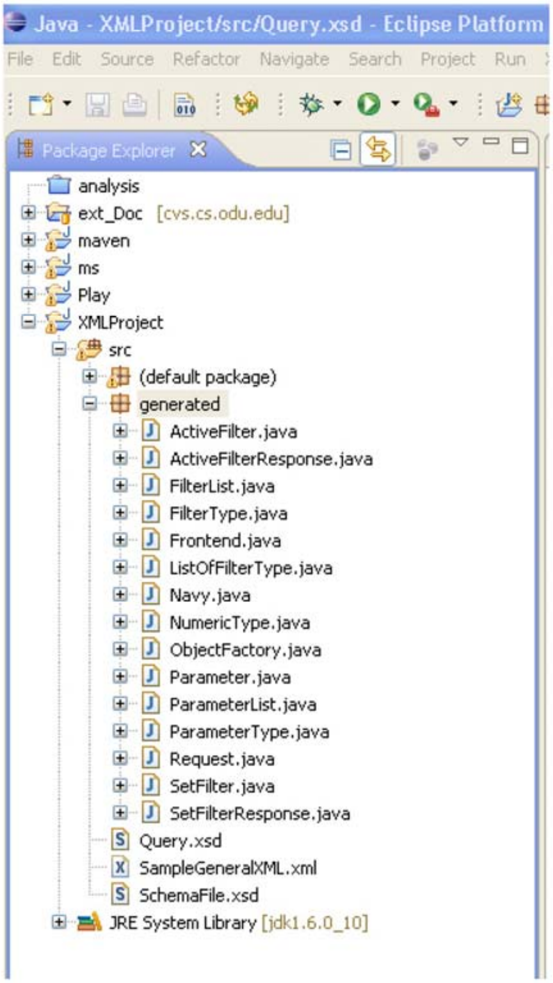


Figure 2: Simplified System Architecture and Filters

The authors utilized the single filter approach (between the front end and back end) in order to demonstrate a proof of concept approach to filtering to ensure the TACOP requirements could be met. The communication is carried through xml messages being exchanged between the frontend and the backend, where we have used marshalling and unmarshalling of xml. To briefly describe XML concepts, we have to know about XML data binding. XML data binding refers to the process of representing the information in an XML document as an object in computer memory.

This allows applications to access the data in the XML from the object rather than using the DOM or SAX to retrieve the data from a direct representation of the XML itself. An XML data binder accomplishes this by automatically creating a mapping between elements of the XML schema of the document we wish to bind and members of a class to be represented in memory. When this process is applied to convert a XML document to an object, it is called unmarshalling. The reverse process, to serialize an object as XML, is called marshalling. After everything is done as described above you should get the hierarchy in eclipse (which we are using) as shown in Figure 3.


Java - XMLProject/src/Query.xsd - Eclipse Platform
File Edit Source Refactor Navigate Search Project Run
Package Explorer

```
analysis
ext_Doc [cvs.cs.odu.edu]
maven
ms
Play
XMLProject
  src
    (default package)
    generated
      ActiveFilter.java
      ActiveFilterResponse.java
      FilterList.java
      FilterType.java
      Frontend.java
      ListOfFilterType.java
      Navy.java
      NumericType.java
      ObjectFactory.java
      Parameter.java
      ParameterList.java
      ParameterType.java
      Request.java
      SetFilter.java
      SetFilterResponse.java
    Query.xsd
    SampleGeneralXML.xml
    SchemaFile.xsd
  JRE System Library [jdk1.6.0_10]
```

Figure 3: Hierarchy After Generating the Classes Using XJC

The next step was to develop a set of use cases between the frontend and backend, such as the frontend making queries like active filters, specific filter parameters, values and the comparators to filter the entities that the user is looking for. A state machine tool was then developed using JAXB and eclipse based on the developed use cases.

The advantage of this approach in comparison with DDM is that the recommended solution is fully and dynamically configurable during runtime. In addition, predefined solutions can be stored as XML files that can be distributed between various users, supporting the FST user community.

Finally, future work will develop the backend filtering engine using DROOLS (or JBoss Rules which is a Java based rule engine) (see Browne, 2009). DROOLS will be used to compare facts about objects in the simulation against user configured filter rules which are added to a Knowledge Based Session.

During follow-on work the proposed three system architecture will also be revisited in efforts to design a TACOP which functions in a less distributed, API-like format.

4.0 CONCLUSION(S)

The paper discusses the requirements and development of an architecture and associated prototype for the TACOP. The prototype provided for the communication between IOS and filtering framework where the communication is carried out through XML messages. With this prototype we have proposed a solution to select from the thousands of simulated entities to display only those entities which fit user requirements, helping to alleviate a common problem experienced by trainers during LVC training events.

5.0 REFERENCES

1) Bizub, W., D. Bryan, E. Harvey (2006). The Joint Live Virtual Constructive Data Translator Framework – Interoperability for a Seamless Joint Training Environment. *Proceedings of the 2006 NATO RTO MSG Symposium, Report RTO-MP-MSG-045*, Rome, Italy, October, paper MP-MSG-045-09-USA

2) Browne, Paul (2009). *JBoss Drools Business Rules* (1st ed.), Packt Publishing, pp. 304, ISBN 1847196063.

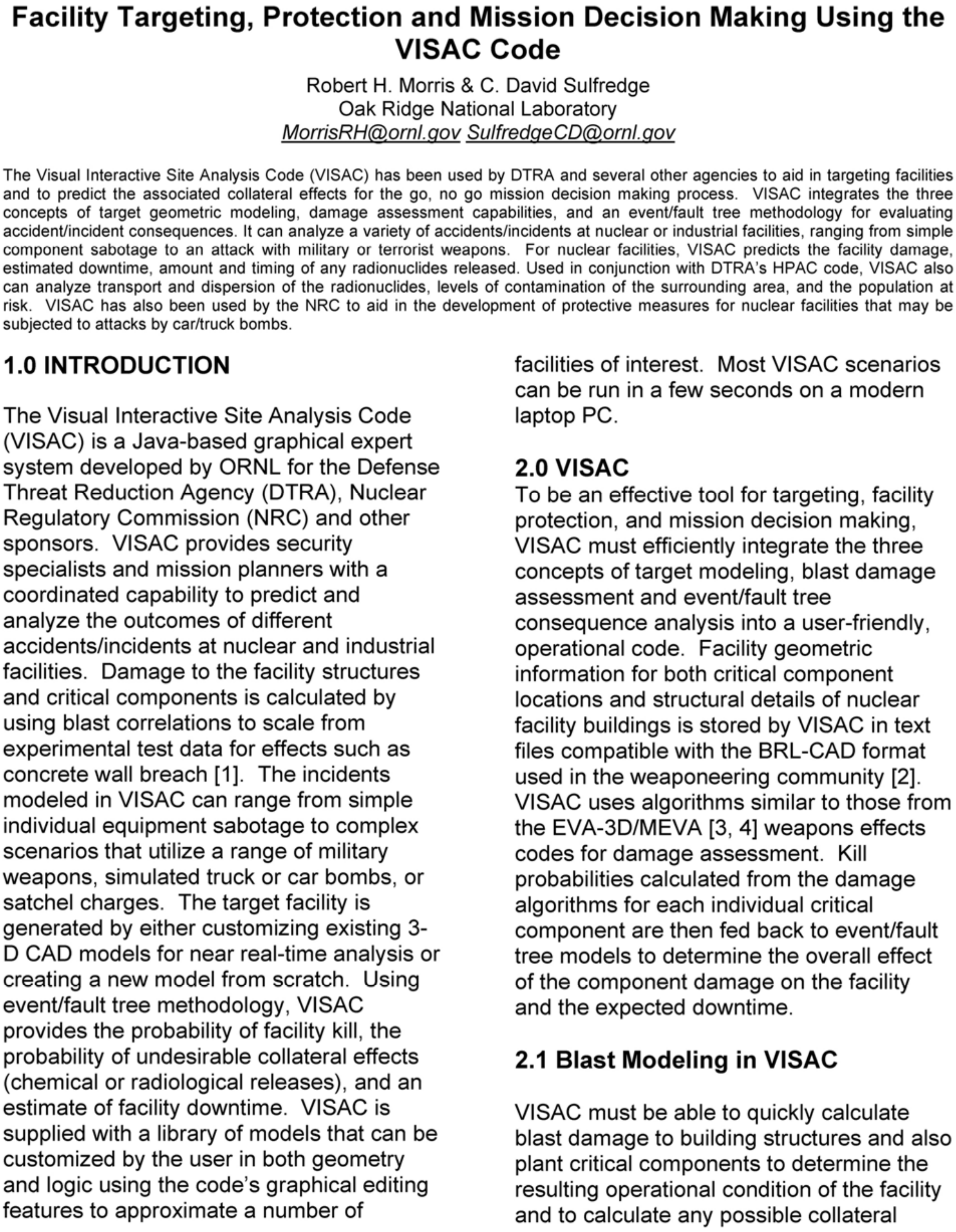
3) IEEE (2010). *1516-2010 IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)-Framework and Rules.*

4) Oracle (2003, Feb.). *Java Architecture fox XML Binding (JAXB) Bindings Schema for JAXB*. Retrieved from http://java.sun.com/xml/ns/jaxb/.

5) Morse, K.L., R. Brunton, J.M. Pullen, P. McAndrews, A. Tolk, J.A. Muguira (2004). An Architecture for Web Services Based Interest Management in Real Time Distributed Simulation, *Proceedings of the 8th IEEE International Symposium on Distributed Simulation and Real Time Applications*, pp. 108-115, October 21-23, Budapest, Hungary.

6) S. Singhal and M. Zyda (1999). *Networked Virtual Environments: Design and Implementation*, ACM Press.

1.4 Facility Targeting, Protection and Mission Decision Making Using the VISAC Code



Facility Targeting, Protection and Mission Decision Making Using the VISAC Code

Robert H. Morris & C. David Sulfredge
Oak Ridge National Laboratory
MorrisRH@ornl.gov SulfredgeCD@ornl.gov

The Visual Interactive Site Analysis Code (VISAC) has been used by DTRA and several other agencies to aid in targeting facilities and to predict the associated collateral effects for the go, no go mission decision making process. VISAC integrates the three concepts of target geometric modeling, damage assessment capabilities, and an event/fault tree methodology for evaluating accident/incident consequences. It can analyze a variety of accidents/incidents at nuclear or industrial facilities, ranging from simple component sabotage to an attack with military or terrorist weapons. For nuclear facilities, VISAC predicts the facility damage, estimated downtime, amount and timing of any radionuclides released. Used in conjunction with DTRA's HPAC code, VISAC also can analyze transport and dispersion of the radionuclides, levels of contamination of the surrounding area, and the population at risk. VISAC has also been used by the NRC to aid in the development of protective measures for nuclear facilities that may be subjected to attacks by car/truck bombs.

1.0 INTRODUCTION

The Visual Interactive Site Analysis Code (VISAC) is a Java-based graphical expert system developed by ORNL for the Defense Threat Reduction Agency (DTRA), Nuclear Regulatory Commission (NRC) and other sponsors. VISAC provides security specialists and mission planners with a coordinated capability to predict and analyze the outcomes of different accidents/incidents at nuclear and industrial facilities. Damage to the facility structures and critical components is calculated by using blast correlations to scale from experimental test data for effects such as concrete wall breach [1]. The incidents modeled in VISAC can range from simple individual equipment sabotage to complex scenarios that utilize a range of military weapons, simulated truck or car bombs, or satchel charges. The target facility is generated by either customizing existing 3-D CAD models for near real-time analysis or creating a new model from scratch. Using event/fault tree methodology, VISAC provides the probability of facility kill, the probability of undesirable collateral effects (chemical or radiological releases), and an estimate of facility downtime. VISAC is supplied with a library of models that can be customized by the user in both geometry and logic using the code's graphical editing features to approximate a number of facilities of interest. Most VISAC scenarios can be run in a few seconds on a modern laptop PC.

2.0 VISAC

To be an effective tool for targeting, facility protection, and mission decision making, VISAC must efficiently integrate the three concepts of target modeling, blast damage assessment and event/fault tree consequence analysis into a user-friendly, operational code. Facility geometric information for both critical component locations and structural details of nuclear facility buildings is stored by VISAC in text files compatible with the BRL-CAD format used in the weaponeering community [2]. VISAC uses algorithms similar to those from the EVA-3D/MEVA [3, 4] weapons effects codes for damage assessment. Kill probabilities calculated from the damage algorithms for each individual critical component are then fed back to event/fault tree models to determine the overall effect of the component damage on the facility and the expected downtime.

2.1 Blast Modeling in VISAC

VISAC must be able to quickly calculate blast damage to building structures and also plant critical components to determine the resulting operational condition of the facility and to calculate any possible collateral effects. There are two common approaches for numerical modeling of blast effects: One method involves hydrocodes such as CTH and DYNA-3D, which are based on first-principle solutions for the conservation equations of mass, momentum, and energy in the shock wave interactions, combined with sophisticated equations of state for the materials involved. While hydrocode calculations are excellent for looking at the details of specific shock wave behavior, the computer run times required are too long to use them for calculations involving large numbers of components or structures. The alternative approach makes use of empirical correlations derived from experimental test data to represent blast effects on components and structures. This second approach is the blast modeling methodology selected for incorporation into VISAC. Using correlations allows VISAC to analyze overall facility vulnerability in a fast-running code without trying to reduce the calculations to first principles. Correlations can generate remarkably accurate results for blast effects as long as they are not extrapolated outside the range of applicable data.

The blast assessment algorithms programmed into VISAC were adapted from correlations that actually date back to empirical test data on weapons effects that was generated by the National Defense Research Committee (NDRC) during World War II [1]. Curve fits to the NDRC test data are available for concrete wall breach probability, blast overpressure as a function of scaled distance from an explosion, and expected overpressure enhancement due to reflective surfaces surrounding the charge.

Once VISAC has applied the NDRC correlations to determine which concrete walls are breached by a blast and how the shock overpressure is propagated through air spaces exposed to the explosion, it is necessary to calculate kill probabilities for the critical components. These component failure probabilities then serve as input for the event/fault tree models that assess overall facility kill probability and the potential for a radiological release from the facility. Thus each critical component in the VISAC facility model must have an associated fragility function expressed in terms of blast overpressure.

An overpressure fragility function consists of a plot showing the component kill probability versus the peak overpressure experienced by the component, as seen in the example function given by Fig. 1. Typically, fragility functions are defined by specifying a minimum overpressure, P_{low}, below which the component kill probability is zero and a maximum overpressure, P_{high}, above which it is unity. The fragility function is then interpolated between P_{low} and P_{high} using either a linear approximation or a logarithmic fit so that any component exposed to an overpressure P between P_{low} and P_{high} is assigned a fractional kill probability between 0 and 1. Estimates of overpressure fragilities for various categories of equipment can be found in Young, et al. [3], Glasstone [5], and Stephens [6]. These sources were used to develop the fragility functions for blast modeling in VISAC.

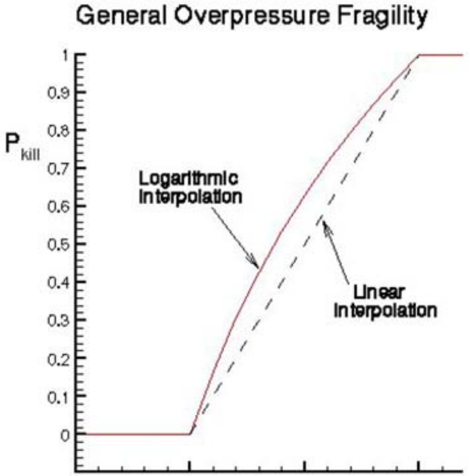


Fig. 1. Example showing the general form for a component fragility function in terms of peak blast overpressure

2.2 Event/Fault Tree Analysis in VISAC

Event tree/fault tree methodology has been applied for decades in the nuclear industry for consequence assessment and determining the probability of reactor core damage during accident sequences [7]. Basically, event trees track the progress of an accident sequence and define the safety systems that can be applied to avoid undesirable consequences. At each branch of an event tree, an underlying fault tree shows the critical components (connected by AND or OR logic gates) necessary for that safety system to function. VISAC adopts this modeling approach to calculate the probability of collateral damage and potential radiological releases associated with attack scenarios at nuclear facilities, and adds another "facility kill" fault tree to determine the probability that the plant would be forced to shut down by the incident.

Ordinary event/fault tree calculations typically involve very small component failure probabilities for the cut sets (groups of components whose simultaneous failure would lead to failure for an overall system fault tree), so that certain mathematical shortcuts such as the rare events approximation [8] or the Esary-Proschan approximation [9] are useful for evaluating sequence probabilities. On the other hand, in vulnerability analysis critical component failure probabilities tend to be high, so these approximations are no longer valid. Therefore VISAC had to introduce some unusual evaluation techniques based on Bayesian analysis and Monte Carlo methods to solve for the sequence probabilities efficiently [10]. VISAC allows a user to set up the system logic model on a detailed basis using an arbitrarily large number of event trees and fault trees referenced to basic events that are critical components in the facility model. A coarser "building level" analysis is also possible, in which only the facility buildings are basic events in the logic model and failure of a building implies loss of function for all critical components and systems located in that building.

2.3 VISAC Facility Downtime Calculation

The cut sets of VISAC's fault tree for facility kill are also vital for estimating facility downtime. Each cut set of the facility kill fault tree consists of a group of equipment without which the plant cannot operate. The overall expected downtime from an attack scenario is thus given by assigning a kill probability and a downtime to each cut set kill path and summing over the cut sets where the overall expected downtime has been normalized by the overall facility kill probability to make it conditional on facility kill. VISAC calculates a strictly serial repair process where damaged components are expected to be fixed sequentially one after the other and also a parallel downtime that assumes workers at the plant can repair the equipment associated with multiple cut sets of the facility kill tree simultaneously. VISAC provides the user both of these downtimes for each scenario.

3.0 VISAC ANALYSIS

VISAC provides a very fast "what if" analysis for various accident/incident scenarios and also facility damage/consequence contour maps. Specifically VISAC returns the facility kill probability, core melt probability for a nuclear reactor, chemical release and fire effects probability for a reprocessing facility and estimates of the downtime for each facility. When used in conjunction with DTRA's HPAC code, VISAC will also provide the transport and deposition of the released material as well as the population at risk.

Incidents include:

- Weapons – satchel charge, truck bomb, military weapon, etc. placed anywhere in or around the facility
- Defined Accident - failure of a specific system
- Region Damage – all components in a specific area fail
- Component damage – individual components fail with specific probabilities.

In addition, multiple incidents can be specified in any combination and as functions of time.

Figure 2 shows a view of a typical VISAC model of a generic nuclear reactor. The critical components are shown in red both inside and outside the facility.

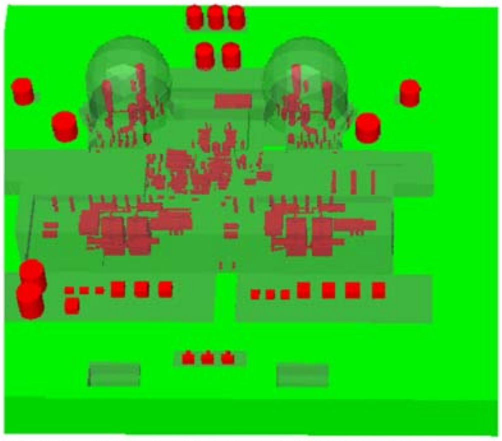


Fig. 2. View of typical VISAC model.

Output from a VISAC scenario can be obtained in tabular or graphical form. For the tabular form, VISAC provides

- Failure probabilities for each fault tree system
- Consequence probabilities for each event tree
- Facility downtime
- Release estimates
- Transport and deposition of released material
- Population at risk
- List of walls broken and hole size estimates

For the graphical form, VISAC can provide views of the facility highlighting the holes in walls and damaged equipment, views of the fault trees showing the broken components and a graphical representation showing the most probable path through the event tree.

Figure 3 shows a typical fault tree with the damaged components crossed out.

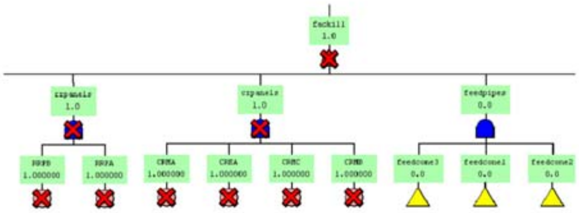


Fig. 3. Typical fault tree showing damaged components from a VISAC scenario.

While running an individual VISAC scenario is very quick, many different locations must be checked to determine the optimum target point for achieving the desired results. Usually when targeting a facility like a nuclear reactor it is desired to maximize the kill probability and the facility down time while minimizing the adverse collateral effects. A secondary consideration to this would be to minimize the population at risk. To achieve this VISAC can be run in a batch mode with a grid of points for where weapons are to be detonated. This code option produces a contour map showing the results at thousands of locations for each particular weapon. Separate contour maps for facility kill probability, probability of collateral damage, and facility downtime can be produced. Grids locations can be inside or outside of the facility. Figures 4 through 6 show the graphical results for facility kill, probability of collateral damage, and downtime from running VISAC on an internal grid for a generic nuclear reactor facility. In the figures red is the highest probability at 80 to 100% and blue is the lowest at less than 20%.

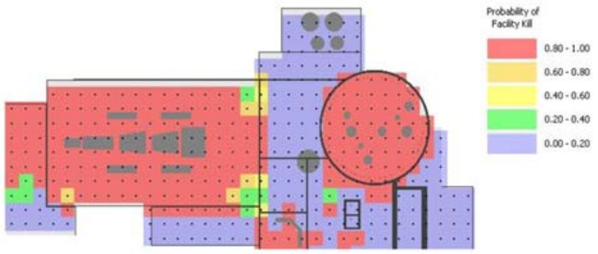


Fig. 4. Facility kill probability for a facility.

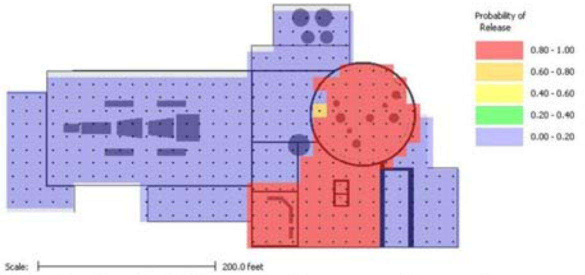


Fig. 5. Probability of adverse collateral damage.

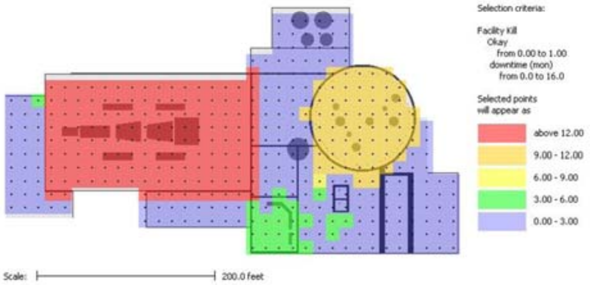


Fig. 6. Expected facility downtime.

By examining these three figures together, a set of locations for the most desirable strike points may be determined to achieve the desired results. The optimum locations determined from the figures can then be used to assess whether the results from the strike are worth the possible unintended collateral damage. This information can then be used as one of the inputs to the mission go, no go decision.

The grid feature can also be used to address protective measures that need to be taken to safeguard a facility. For example, to determine the measures that need to be taken to safeguard a facility from a truck bomb, grids can be outside the facility and a truck bomb detonated at the grid locations. Figure 7 shows the results from this type of calculation. As can be seen from the figure the risk to the facility from vehicle bombs can be greatly reduced by improving the security at only a few locations.

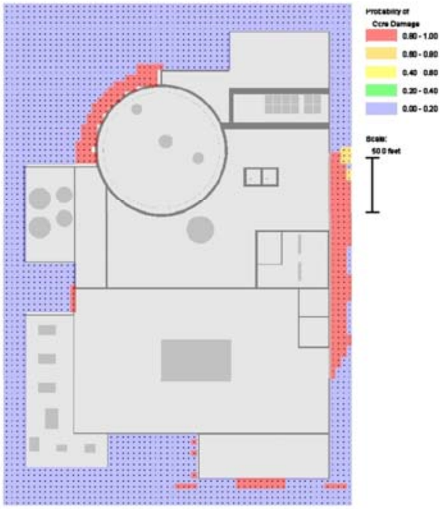


Fig. 7. Facility protective measure for defense against a truck bomb.

As mentioned earlier, if the analyst has a copy of DTRA's HPAC 5 software installed on their PC, then VISAC can provide transport and deposition of any releases as well as the population at risk. These plots are available in the traditional HPAC format or can be overlaid on Google maps. Figures 8 and 9 are examples of these types of plots.

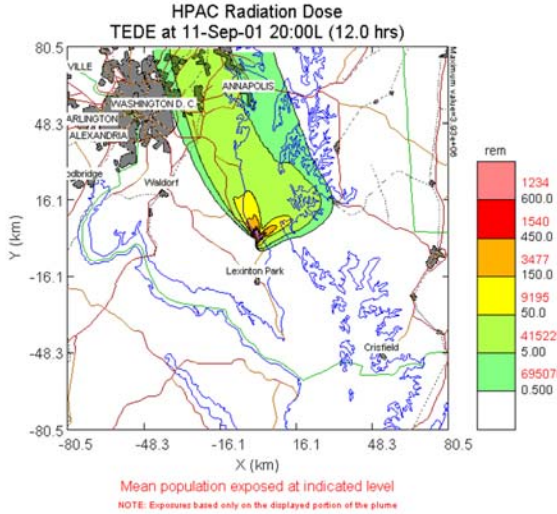


Fig. 8. Typical HPAC plot for radionuclide

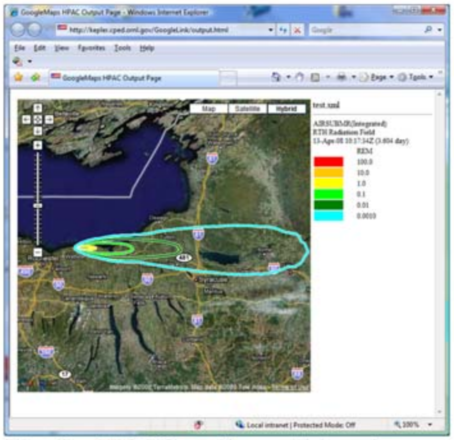

GoogleMaps HPAC Output Page - Windows Internet Explorer
http://kepler.cped.ornl.gov/GoogleLink/output.html
Google
File Edit View Favorites Tools Help
GoogleMaps HPAC Output Page
Page
Tools
Map
Satellite
Hybrid
test.xml
AIRSUBMR(Integrated)
RTH Radiation Field
13-Apr-08 10:17:34Z (3.604 day)
REM
100.0
10.0
1.0
0.1
0.01
0.0010
Local intranet | Protected Mode: Off
100%

Fig. 9. VISAC radionuclide release overlaid on GoogleMaps.

4.0 CONCLUSIONS

- VISAC integrates target geometric modeling, blast damage assessment, and event/fault tree consequence analysis.
- The VISAC code successfully uses correlation blast modeling to generate quick facility vulnerability results.
- Results from the VISAC analysis can be used to
  - Optimize target locations to achieve the desired results
  - Provide the locations requiring increased protective measure
  - Aid in the mission go no-go decision making

5.0 REFERENCES

1. M. P. White, et al. *Effects of Impact and Explosion,* Summary Technical Report of Division 2, NDRC, Washington, DC, 1946.
2. U. S. Army Ballistic Research Laboratory, *The Ballistic Research Laboratory CAD Package, Release 4.0*, Aberdeen, MD, December 1991.
3. L. A. Young, B. K. Streit, K. J. Peterson, D. L. Read, F. A. Maestas, *Effectiveness/Vulnerability Assessments in Three Dimensions (EVA-3D) Versions 4.1F and 4.1C User's Manual – Revision A*, WL-TR-96-7000, Applied Research Associates, Inc., Albuquerque, NM, November 29, 1995.
4. P. E. Dunn, J. E. Madrigal, D. A. Parsons, J. C. Partch, D. A. Verner, and L. A. Young, *Modular Effectiveness/ Vulnerability Assessment (MEVA) Software User's Manual*, Applied Research Associates, Inc., Albuquerque, NM, April 23, 1999.
5. S. Glasstone, *The Effects of Nuclear Weapons*, United States Atomic Energy Commission, Washington, DC, 1962.
6. M. M. Stephens, *Minimizing Damage to Refineries from Nuclear Attack, Natural and Other Disasters*, U. S. Department of the Interior, Office of Oil and Gas, Washington, DC, 1970.
7. U. S. NRC, *Reactor Safety Study-An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants, Report WASH-1400*, NUREG-75/014, 1975.
8. McCormick, N. J., *Reliability and Risk Analysis*, Academic Press, New York, 1981.
9. Isograph Ltd., *FaultTree+ Software*

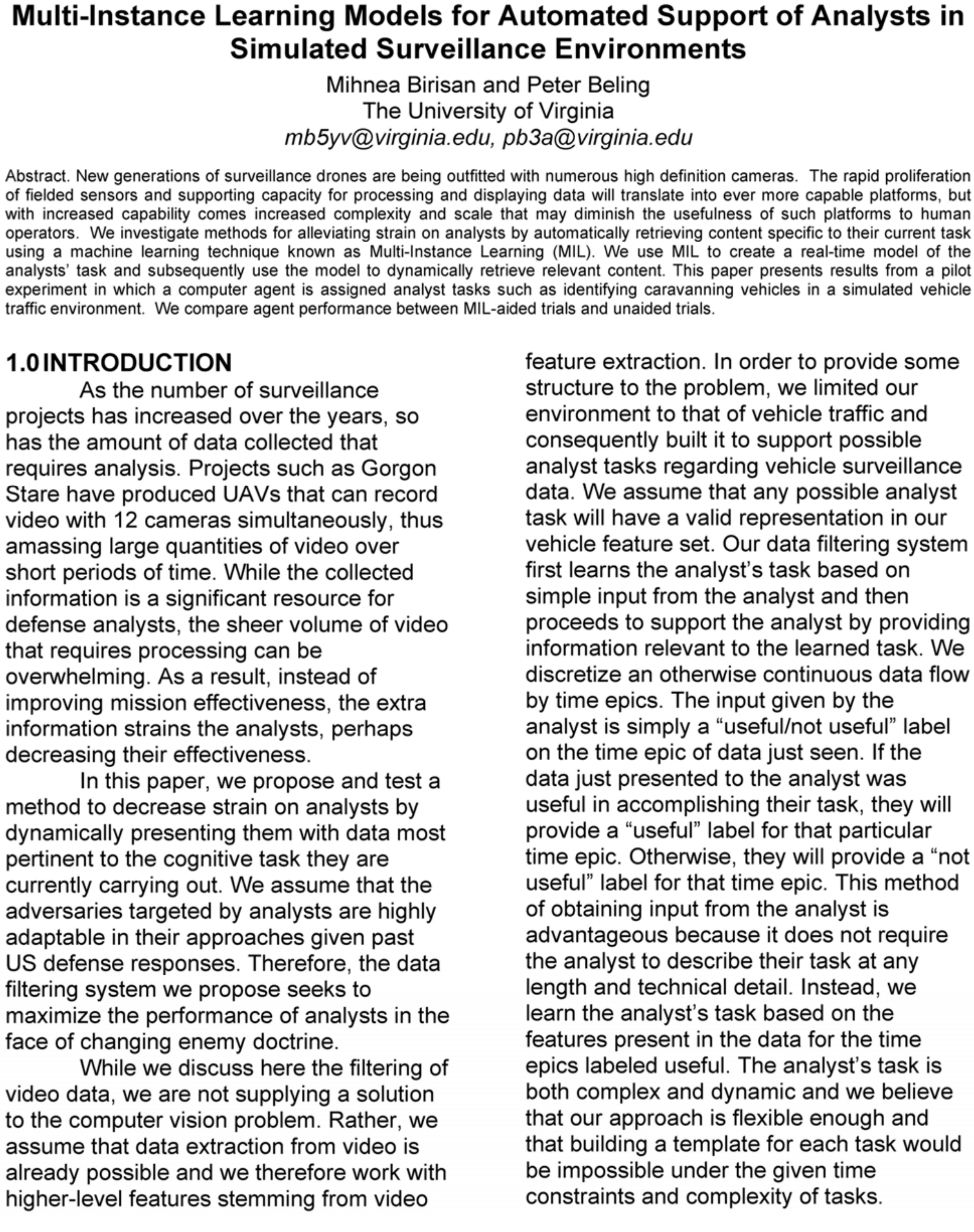
Documentation, 1996.

10. Peplow, D. E., Sulfredge, C. D., Sanders, R. L., Morris, R. H., *Calculating Nuclear Power Plant Vulnerability Using Integrated Geometry and Event/Fault-Tree Models,* Nuclear Science and Engineering 146, 71-87 (2004).

6.0 ACKNOWLEDGMENT

Notice: This manuscript has been authored by UT-Battelle, under Contract No. DE-AC0500OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

1.5 Multi-Instance Learning Models for Automated Support of Analysts in Simulated Surveillance Environments



Multi-Instance Learning Models for Automated Support of Analysts in Simulated Surveillance Environments

Mihnea Birisan and Peter Beling
The University of Virginia
mb5yv@virginia.edu, pb3a@virginia.edu

Abstract. New generations of surveillance drones are being outfitted with numerous high definition cameras. The rapid proliferation of fielded sensors and supporting capacity for processing and displaying data will translate into ever more capable platforms, but with increased capability comes increased complexity and scale that may diminish the usefulness of such platforms to human operators. We investigate methods for alleviating strain on analysts by automatically retrieving content specific to their current task using a machine learning technique known as Multi-Instance Learning (MIL). We use MIL to create a real-time model of the analysts' task and subsequently use the model to dynamically retrieve relevant content. This paper presents results from a pilot experiment in which a computer agent is assigned analyst tasks such as identifying caravanning vehicles in a simulated vehicle traffic environment. We compare agent performance between MIL-aided trials and unaided trials.

1.0 INTRODUCTION

As the number of surveillance projects has increased over the years, so has the amount of data collected that requires analysis. Projects such as Gorgon Stare have produced UAVs that can record video with 12 cameras simultaneously, thus amassing large quantities of video over short periods of time. While the collected information is a significant resource for defense analysts, the sheer volume of video that requires processing can be overwhelming. As a result, instead of improving mission effectiveness, the extra information strains the analysts, perhaps decreasing their effectiveness.

In this paper, we propose and test a method to decrease strain on analysts by dynamically presenting them with data most pertinent to the cognitive task they are currently carrying out. We assume that the adversaries targeted by analysts are highly adaptable in their approaches given past US defense responses. Therefore, the data filtering system we propose seeks to maximize the performance of analysts in the face of changing enemy doctrine.

While we discuss here the filtering of video data, we are not supplying a solution to the computer vision problem. Rather, we assume that data extraction from video is already possible and we therefore work with higher-level features stemming from video feature extraction. In order to provide some structure to the problem, we limited our environment to that of vehicle traffic and consequently built it to support possible analyst tasks regarding vehicle surveillance data. We assume that any possible analyst task will have a valid representation in our vehicle feature set. Our data filtering system first learns the analyst's task based on simple input from the analyst and then proceeds to support the analyst by providing information relevant to the learned task. We discretize an otherwise continuous data flow by time epics. The input given by the analyst is simply a "useful/not useful" label on the time epic of data just seen. If the data just presented to the analyst was useful in accomplishing their task, they will provide a "useful" label for that particular time epic. Otherwise, they will provide a "not useful" label for that time epic. This method of obtaining input from the analyst is advantageous because it does not require the analyst to describe their task at any length and technical detail. Instead, we learn the analyst's task based on the features present in the data for the time epics labeled useful. The analyst's task is both complex and dynamic and we believe that our approach is flexible enough and that building a template for each task would be impossible under the given time constraints and complexity of tasks.

Driving our data filtering system is a machine-learning algorithm know as Multi-Instance Learning (MIL). MIL is useful given our problem because it is responsive to changes in the analyst's task and because it does not require a label for every piece of data. A detailed description of the architecture and functioning of MIL follows in a later section.

Finally, in order to evaluate the value added to mission effectiveness by the MIL-driven data filtering system, we devised a way to test analyst mission effectiveness both with the MIL filtering system in place and without it. To this end, we devised a series of tasks relating to vehicle information and assigned them to a computer agent, which performed them both with and without MIL aid in a simulated environment. The agent was scored under both aid conditions and the scores were compared across all tasks. We will show that the agent performed better with MIL aid.

2.0 BACKGROUND

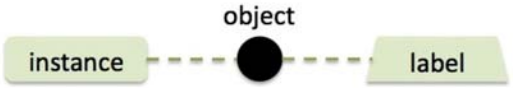
Machine learning can broadly be divided into two different approaches: *supervised learning* and *unsupervised learning*. In the supervised learning approach, the learning algorithm is provided with a label for every training example. Oftentimes, it is not feasible or possible to provide labels for training examples. Thus, in unsupervised learning, the learning algorithm is provided with completely unlabeled training examples, with learning algorithms in unsupervised learning stemming from clustering principles. MIL blurs the difference between supervised and unsupervised learning because it use partially labeled training examples.

MIL was first introduced in the context of the drug activity prediction problem described in reference [2]. Reference [2] also proposed the first algorithm to solve the MIL problem. In drug activity prediction, one must predict if a given molecule will bind to a target binding site. The binding site is located on a much larger molecule, such as a protein, and has a very specific shape, making it impossible for any given molecule to bind there unless is has the perfect matching shape. Incomplete information stems from the fact that while it is possible to tell whether a molecule did or did not bind to the target site, it is impossible to tell what shape it had when it did bind to the target site. This happens because each molecule can take on several different shapes based on its bond angles. Thus, the positive label given to a molecule that did bind to the target site is ambiguous in that it does not define the shape that the molecule took on when the binding occurred.

Following reference [2], reference [3] tested a new MIL algorithm on the drug activity data set and also tried two new applications: forming a concept of what a person looks like from a series of labeled pictures and dealing with noise in stock selection. Reference [4] then used the new algorithm from reference [3] for natural scene image classification. Reference [6] used the drug activity data set to test yet another MIL algorithm. Two more applications of MIL have been in automated, content-based image retrieval described in reference [7] and text categorization describe in reference [1]. Reference [7] used the same algorithm as reference [3], but applied it to a different problem (that of content-based image retrieval), thus showing that MIL algorithms are flexible enough to have a variety of applications as long as the structure of the problem is maintained. The content-based image retrieval problem, specifically, was also worked on by reference [5], who proposed a new algorithm for this problem. In this paper we use the algorithm first described in reference [3].

3.0 THE ARCHITECTURE OF MIL

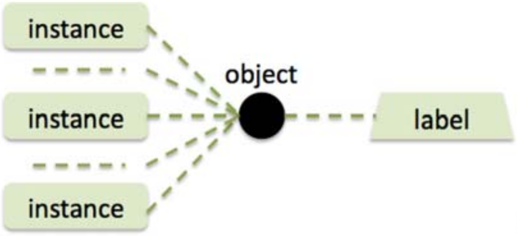
Before we proceed, we must establish some terminology. We will refer to examples, such as the training examples discussed in the Background section, as *objects*. Each object has a representation in feature space known as an *instance*. In the supervised learning case, each object is described by a single instance and each instance has a *label*. Figure 1 shows the one-to-one relationship of instances, objects, and labels.



#

Figure 1 – Supervised Learning Labels

In the MIL case, however, there is no one-to-one mapping of instances, objects, and labels. Instead, each object can be represented by multiple instances in feature space. In the MIL case, labels are assigned to each object, not to each instance. Figure 2 depicts the architecture of labels in MIL.



#

Figure 2 – MIL Labels

In MIL, the term *bag* is used to refer to an object. The term bag is used in order to illustrate that an object can “contain” or be described by several instances – a bag of instances. Since the label is not placed on each instance, but rather on the bag as a whole, rules must be set for labeling a bag as a function of the instances. A bag is labeled *positive* if at least one instance in the bag is positive. A bag is labeled *negative* if all the instances in the bag are negative. When looking at a bag labeled positive, it is ambiguous which instance triggered the positive label. MIL algorithms examine the instances in positive bags in order to find a feature space representation of the instances that triggered positive labels.

4.0 SIMULATION

We constructed a vehicle simulation environment to test if MIL-based data filtering would add value to operator mission effectiveness. In the simulation each vehicle exhibits either a normal behavior or a rogue behavior. Normal behaviors consist in entering traffic through an entry point, following traffic rules, and ultimately leaving the simulation through an exit point. Rogue behaviors consist in abnormal patterns that contradict traffic rules. In addition, each vehicle will also have two characteristics: color and type – car or truck. Table 1 shows each type of behavior and the rule used to determine if that behavior applies to a given vehicle.

Table 1 – Vehicle Behavior Definitions

Behavior	Rule
Speeding	Instantaneous speed >> speed limit
Slow Moving	Instantaneous speed << speed limit
Caravanning	2 or more vehicles with max Euclidian distance < epsilon distance, matching speed at every step within some epsilon speed, matching at least 4 turns
Abandoning	In a set of caravanning vehicles, one vehicle stops
Circling	Vehicle makes a series of more than 8 same-direction turns
Multiple U-Turns	Vehicle reverses coordinates

In MIL terminology, vehicle behaviors are instances, time epics are bags, and the description of behaviors in terms of low-level features defines the feature space. To support the instances (vehicle behaviors) in our simulation, we defined a feature space that is consistent with vehicle behavior metrics. Table 2 shows the features present in the simulation.

Table 2 – Feature Set

Feature Name	Type
Speeding Cars Present	Boolean

Number Speeding Cars	Integer
Slow Moving Cars Present	Boolean
Number Slow Moving Cars	Integer
Caravans Present	Boolean
Number Caravans	Integer
Abandoned Cars Present	Boolean
Number Abandoned Cars	Integer
Circling Cars Present	Boolean
Number Circling Cars	Integer
Multiple U-Turn Cars Present	Boolean
Number Multiple U-Turn Cars	Integer
Number Cars	Integer
Number Trucks	Integer
Number Red Vehicles	Integer
Number Blue Vehicles	Integer
Global Maximum Speed	Integer
Global Minimum Speed	Integer

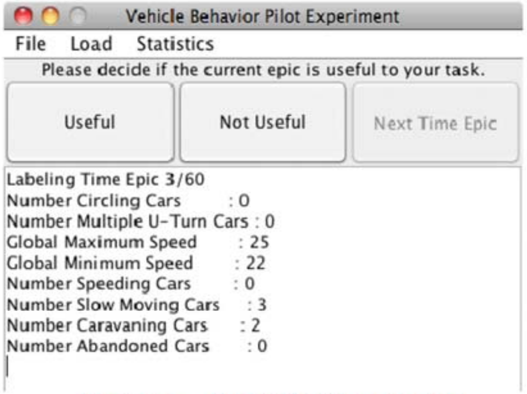
The simulation will be launched in two stages: the pilot stage and the full simulation stage. This paper describes the structure and test results on the pilot stage of the simulation and outlines the structure of the full simulation stage.

4.1 Pilot Stage

The pilot stage simulation is written in Java and is designed to be a proof of concept for MIL-based data filtering in a defense analysis environment. Since the pilot simulation is smaller in scale, it contains a subset of all the features present above. The features present are illustrated in Figure 3 below. The pilot simulation consists of three main components: the GUI, the simulation engine, and the MIL classifier.

4.1.1 Pilot Stage GUI

The pilot stage GUI is designed solely to obtain labels from an analyst or agent on different time epics. Figure 3 shows the pilot simulation GUI. The top part of the GUI allows the experiment organizer to load the simulation for each participant or agent. Each agent can then use the three buttons to provide a label on the summary data for a given time epic and to move on to the next time epic. Below the input buttons is a window that shows the summary data for a given time epic. The agent is only allowed to view the next time epic after they have provided a label for the current epic.



Please decide if the current epic is useful to your task.

```
Labeling Time Epic 3/60
Number Circling Cars       : 0
Number Multiple U-Turn Cars : 0
Global Maximum Speed       : 25
Global Minimum Speed       : 22
Number Speeding Cars       : 0
Number Slow Moving Cars     : 3
Number Caravaning Cars     : 2
Number Abandoned Cars       : 0
```

#

Figure 3 – Pilot Simulation GUI

4.1.2 Pilot Stage Simulation Engine

The simulation engine is in charge of loading the summary data for each time epic, presenting that to the agent, recording the label for each time epic, and finally instantiating the MIL classifier with the labeled data for each time epic. The simulation engine takes as input a text file containing the summary data for 60 training time epics. The engine reads the text file line by line, displays it to the agent, and then records the same data along with the labels provided by the agent in a new text file formatted to be readable to the MIL classifier. Once the agent has finished labeling each of the 60 training time epics according to their task, the simulation engine instantiates and trains the MIL classifier using the labeled data file as input.

Once the classifier object is trained on the labeled data, a new raw data file is presented to the classifier. This new file contains summary data for 60 evaluation time epics. Based on the concept learned in training, the classifier now predicts a label for each of the 60 evaluation time epics. The agent then proceeds to the first of two evaluation phases. The first evaluation phase is an aided phase where the agent is shown only time epics that the MIL classifier labels as matching the learned concept of the agent's task. In the evaluation phase, the "Useful/Not Useful" input buttons no longer provided labels for the classifier, but rather score the classifier, with the classifier's score increasing for every "Useful" time epic label provided by the agent in evaluation. After the aided evaluation phase described above, the agent is shown an equal number of randomly selected time epics. These time epics are not selected by the MIL classifier and thus may or may not be relevant to the agent's task. Once again, a score is incremented each time the agent labels a time epics as "Useful". In the Results section, we compare the evaluation scores in the MIL-aided and the unaided evaluation phases.

4.1.3 Pilot Stage MIL Classifier

The pilot stage MIL classifier is built based on the Diverse Density algorithm introduced by reference [3]. The classifier is instantiated in the pilot simulation using a Java jar file from the WEKA data-mining package. The simulation engine utilizes the classifier's training, prediction, and cross evaluation functions. Following general data mining rules, the simulation engine presents completely different data sets to the classifier for training and prediction. For each task we also run a cross evaluation on the prediction data set. The cross evaluation function uses parts of the prediction data set for training and parts for evaluation, recursively changing which parts are used for training and which for evaluation. We will also present cross evaluation scores across all tasks in the Results section.

4.2 Full Simulation Stage

The full simulation stage will be built on the existing pilot stage. The goal is to design a visual way to present each time epic to a human participant as opposed to a computer agent. Instead of showing the participant summary data for each time epic, the simulation will instead show a map with vehicles moving from entry to exit points. Each vehicle will have an associated track – GPS coordinates for each time step. Most vehicles will have normal, random behaviors, but some vehicles will exhibit "rogue" behaviors. Rogue behaviors will consist in speeding, moving too slowly, two or more cars caravanning, cars being abandoned, and so forth. While each time epic unfolds, the simulation engine will compute the value of each feature in feature space. Instead of labeling the time epic based on parsed summary data, the participants will have to observe how each time epic unfolds by following the cars and seeking behaviors relevant to their task. This method of presenting the participants with information is more realistic and similar to what an analyst would experience in a defense environment.

5.0 EVALUATING MIL

We used an agent-based simulation to measure if the MIL-based data filtering we propose in this paper can add value to analyst mission effectiveness. To that end, we created a list of fourteen tasks relating to vehicle behavior that an analyst might be interested in. Table 3 shows a list of the tasks. Each task is based in identifying at least one rogue vehicle behavior.

Table 3 – Tasks

Task ID	Task Type	Task
1	Simple	Identify ABANDONED vehicles
2	Simple	Identify CARAVANS
3	Simple	Identify CIRCLING vehicles
4	Simple	Identify SLOW vehicles
5	Simple	Identify SPEEDING vehicles
6	Simple	Identify U-TURN vehicles
7	Composite 1/2	Identify ABANDONED & CARAVANS
8	Composite 1/2	Identify CIRCLING & U-TURNS

9	Composite 1/2	Identify SLOW & SPEEDING
10	Composite 1/2	Identify SPEEDING & CARAVANS
11	Composite 2/3	Identify CARAVANS & ABANDONED & SLOW
12	Composite 2/3	Identify SLOW & CIRCLING & U-TURNS
13	Composite 2/3	Identify SLOW & SPEEDING & U-TURNS
14	Composite 2/3	Identify SPEEDING & CARAVANS & U-TURNS

Tasks 1 through 6 ask the agent to provide positive labels for time epics that exhibit a single vehicle behavior given in the task. Tasks 7 through 10 ask the agent to provide positive labels for time epics that exhibit at least one of two vehicle behaviors given in the task. Finally, tasks 11 through 14 ask the agent to provide positive labels for time epics that exhibit at least two of three vehicle behaviors given in the task.

A computer agent was assigned each of the fourteen tasks in turn. The agent labeled all 60 training time epics according to the task at hand. After the training phase, the agent proceeded to the evaluation phase. In the aided evaluation phase, the agent was only shown time epics that the MIL classifier labeled as matching the agent's task. The agent was given a positive point for every positive label it assigned to time epics in the evaluation phase. In the unaided evaluation phase, the agent was shown a random set of time epics and again scored on the number of positive labels it assigned. The evaluation data set did not contain exactly the same number of time epics matching each of the fourteen tasks. To be fair, the agent was shown exactly the same number of random time epics in the unaided phase as the number of time epics matching the agent's task in the aided phase. To illustrate, suppose that in the training phase, the MIL classifier learned that the agent had been assigned task 2. In the aided evaluation, the MIL classifier showed the agent all time epics matching task 2 in the evaluation data set. Suppose there were 13 matches. Then, the agent was also shown 13 random time epics in the unaided evaluation some of which happened to match task 2 and some that did not. This was done to ensure that the agent had a chance to score the same number of points in both the aided and the unaided evaluations.

As we discussed, the next stage in the simulation will provide real human subjects with a visual display of each time epic. This will mimic a defense analysis environment better than the pilot simulation, but will introduce the possibility of human error into all measurements. In order to simulate the effect of human error on the accuracy of the MIL classifier, we ran six more trials with imperfect labels. In these trials, the agent injected 1, 2, or 3 wrong labels when labeling the 60 training time epics. The wrong labels were both false positives (i.e. a bag was incorrectly labeled positive even though it did not match the task at hand) and false negatives (i.e. a bag was labeled negative even though it did match the task at hand), thus adding six more trials in addition to the perfect labels trial. This approach is perhaps unfair toward the MIL classifier because we simulate human error or indecision in the training phase, yet in the evaluation phase, the agent makes no mistakes in labeling, thus adding negative bias to the accuracy of the classifier. Nonetheless, the point was to stress the classifier by simulating real-world conditions. The Results section shows the performance of the MIL classifier across the perfect labels trial as well as the six imperfect labels trials.

6.0 RESULTS

We present the results from all seven trials as lift charts. We graph the scores form the unaided evaluations on the x-axis and the scores from the aided evaluations on the y-axis. If the aided and unaided evaluations exhibited similar performance, all points would lie on the 45-degree line. On the other hand, if the aided evaluation scores are higher than the unaided scores, then we expect the points to lie above the 45-degree line. Figure 4 shows the results from the perfect labels trial.

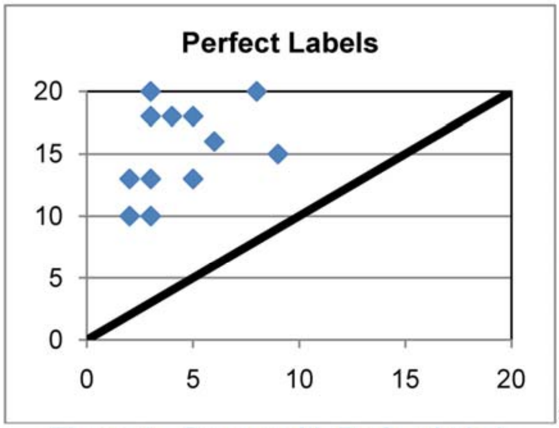


Figure 4 – Scores with Perfect Labels #

The scores from all fourteen tasks are in the top left of the chart, showing that the MIL algorithm has added value to mission effectiveness by showing the agent time epics that matched its task in the aided evaluation phase. Figure 5 shows the results from the two trials with one wrong label. In this and all following figures the green (triangle) points represent false positive labels and the blue (diamond) points represent false negative labels.

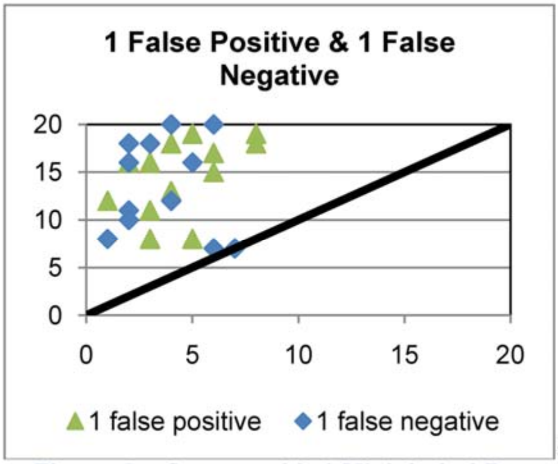


Figure 5 – Scores with 1 Mislabeled Bag #

Here we see that some of the tasks have scores that are closer to the 45-degree line, suggesting less lift from the MIL classifier. Nonetheless, the majority of task scores remain in the top left region of the chart. Figure 6 shows the results with two wrong labels.

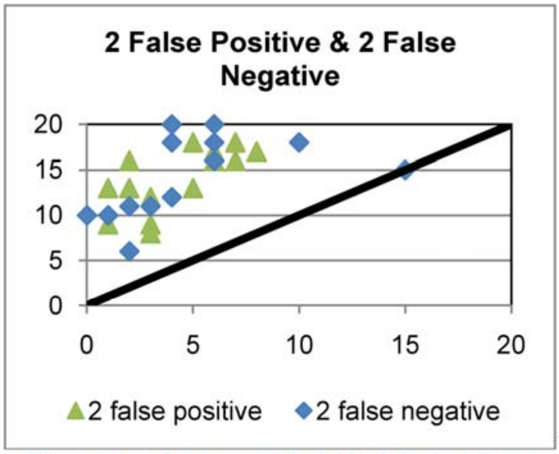


Figure 6 – Scores with 2 Mislabeled Bags #

Finally, Figure 7 shows the results with three wrong labels.

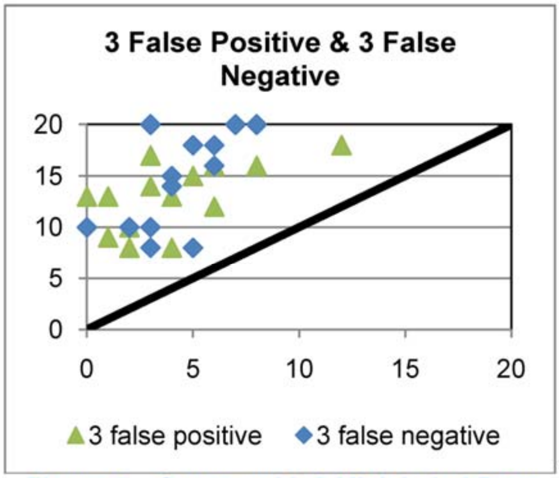
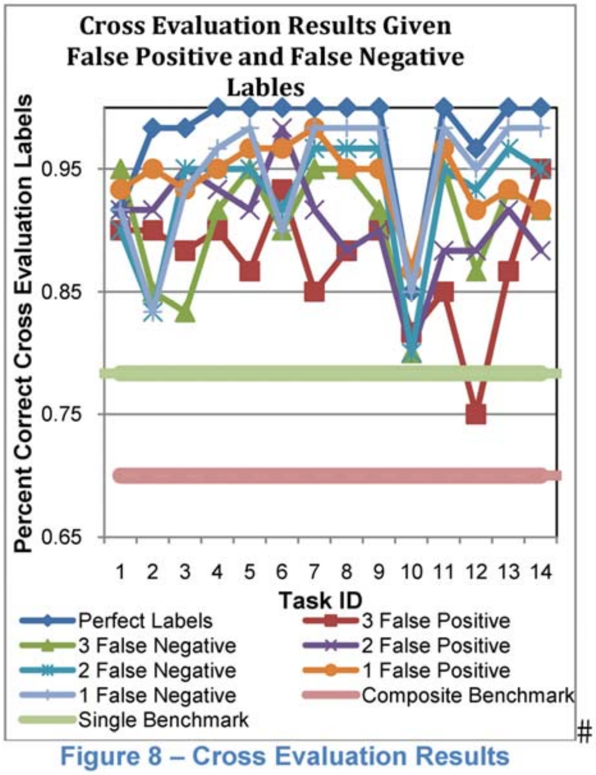


Figure 7 – Scores with 3 Mislabeled Bags #

The difference between the one wrong label case and the two wrong labels case is not immediately noticeable. In fact, it appears that, overall, the one wrong label case resulted in lower scores than the two wrong labels case. In the three wrong labels case it is noticeable that aided scores are overall lower, detracting from the lift of the MIL classifier. Table 4 shows average score values across all tasks by the number of wrong labels. Averaging across all label types, we observe that the agent obtained aided evaluation scores that were 3.3 times higher than unaided evaluation scores.

Table 4 – Average Scores Across Tasks

	Score	
Label Type	Aided	Unaided
Perfect	15.50	4.57
1 False Positive	14.14	4.35
1 False Negative	13.79	3.71
2 False Positive	13.93	4.21
2 False Negative	14.64	4.79
3 False Positive	13.00	4.07
3 False Negative	14.07	4.14



#

Figure 8 – Cross Evaluation Results

Given the current data, it is difficult to conclude if false positive or false negative training labels are more detrimental to the accuracy of the MIL classifier. Thus far, the scores show that false negative labels were more detrimental to the classifier if only one bag was mislabeled. If two or three bags were mislabeled, false positive labels were more detrimental to the accuracy of the classifier.

Figure 8 shows the cross evaluation results obtained on the evaluation data set. Each line shows the percentage of correct labels from the MIL classifier given different numbers of wrong training labels. Naturally, the perfect labels line is highest in the chart, indicating best classifier performance. The green horizontal line marks the single behavior task benchmark. In other words, if the classifier labeled all bags negative, it would still get 78% correctly labeled bags because only 22% of bags have true positive labels in the evaluation data set. The red horizontal line marks the composite behavior task benchmark. The classifier would get 70% correct labels if it labeled all bags negative because there were only 30% true positive bags in the evaluation data set. It is interesting to note that none of the single behavior tasks (1-6) dipped below the single benchmark and none of the composite behavior tasks (7-14) dipped below the composite benchmark, regardless of the number of wrong training labels. It is also noteworthy that all cross evolution results exhibited lower performance on tasks 10 and 12, indicating that there exists possible task dependence in the performance of the MIL classifier. To determine if such dependence exists with statistical significance, it is necessary to add more tasks and trials to the agent-based pilot experiment.

7.0 CONCLUSIONS

Based on the results from the pilot simulation, the Multi-Instance Learning classifier added value to agent mission effectiveness. The robustness of the results in the face of mislabeled training bags suggests that the classifier will continue to add value to mission effectiveness as we transition from agent-based simulations to human subject experiments on the full simulation.

With the current data it is not possible to determine with any statistical significance if false positive or false negative training labels are more detrimental to concept learning on the part of the MIL classifier. However, given more trials on the full simulation it may be possible to determine which type of label is more detrimental. This information is vital in creating a fully featured data filtering system for defense analysis applications.

This proof-of-concept simulation served to show that (1) a data filtering system is useful in relieving strain from today's information-flooded defense analysts and (2) that employing machine learning techniques is a feasible approach in building such a system.

8.0 REFERENCES

[1] S. Andrews, I. Tsochantaridis, T. Hofmann. Support Vector Machines for Multiple-Instance Learning. *NIPS 2002.*

[2] T. G. Dietterich, R. H. Lathrop, T. Lozano-Perez. Solving the Multiple-Instance Problem with Axis-Parallel Rectangles. *Artificial Intelligence Journal,* 89, 1997.

[3] O. Maron , T. Lozano-Pérez, A framework for multiple-instance learning, *Proc. of the 1997 Conf. on Advances in Neural Information Processing Systems* 10, p.570-576, 1998.

[4] O. Maron & A. L. Ratan, (1998). Multiple-instance learning for natural scene classification. In *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 341–349).

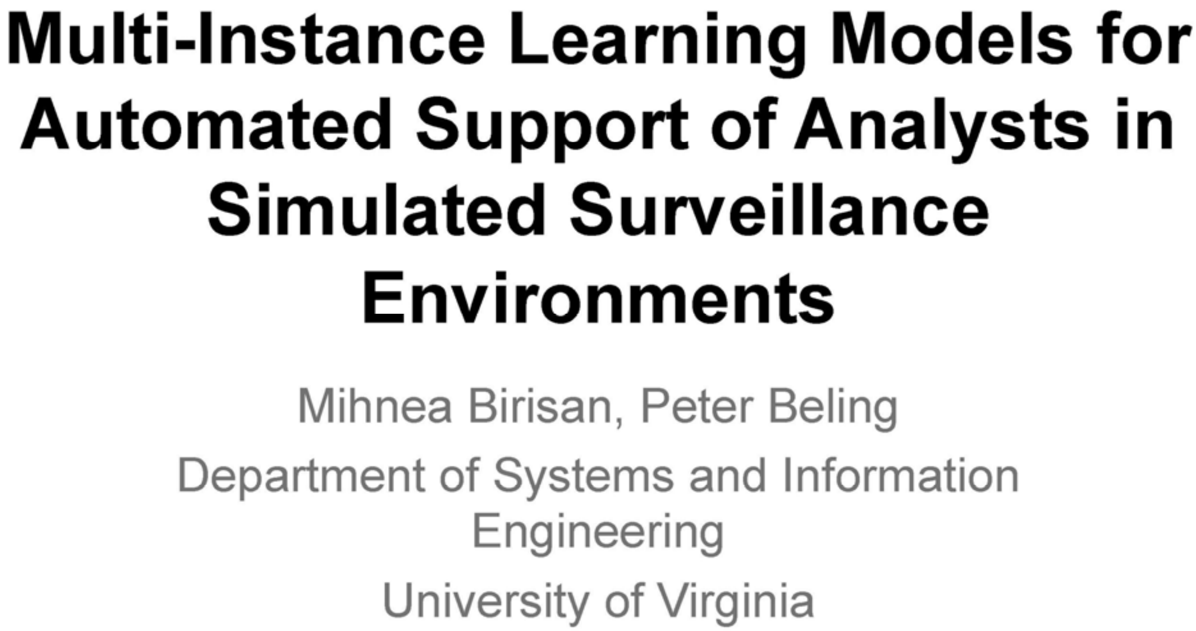
[5] S. Tong & E. Chang, (2001). Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia* (pp. 107–118).

[6] J. Wang and J. D. Zucker. Solving the multiple-instance problem: a lazy learning approach. *Proc. 17th Int'l Conf. on Machine Learning*, pp. 1119-1125, 2000.

[7] C. Yang and T. Lozano-Perez. Image database retrieval with multiple-instance learning techniques. *Proc. of the 16th Int. Conf. on Data Engineering*, pp.233-243, 2000.

ACKNOWLEDGMENT

We would like to thank Brian Rigling of SAIC for supporting Mihnea Birisan in this research.

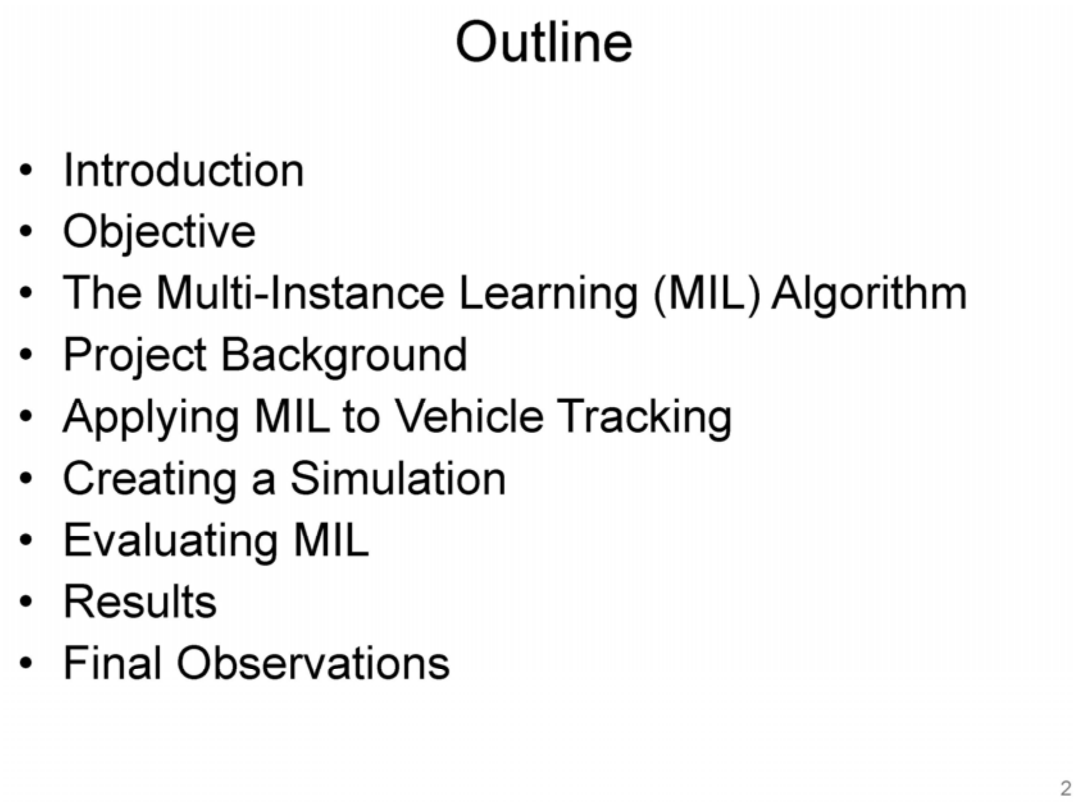


Multi-Instance Learning Models for Automated Support of Analysts in Simulated Surveillance Environments

Mihnea Birisan, Peter Beling

Department of Systems and Information Engineering

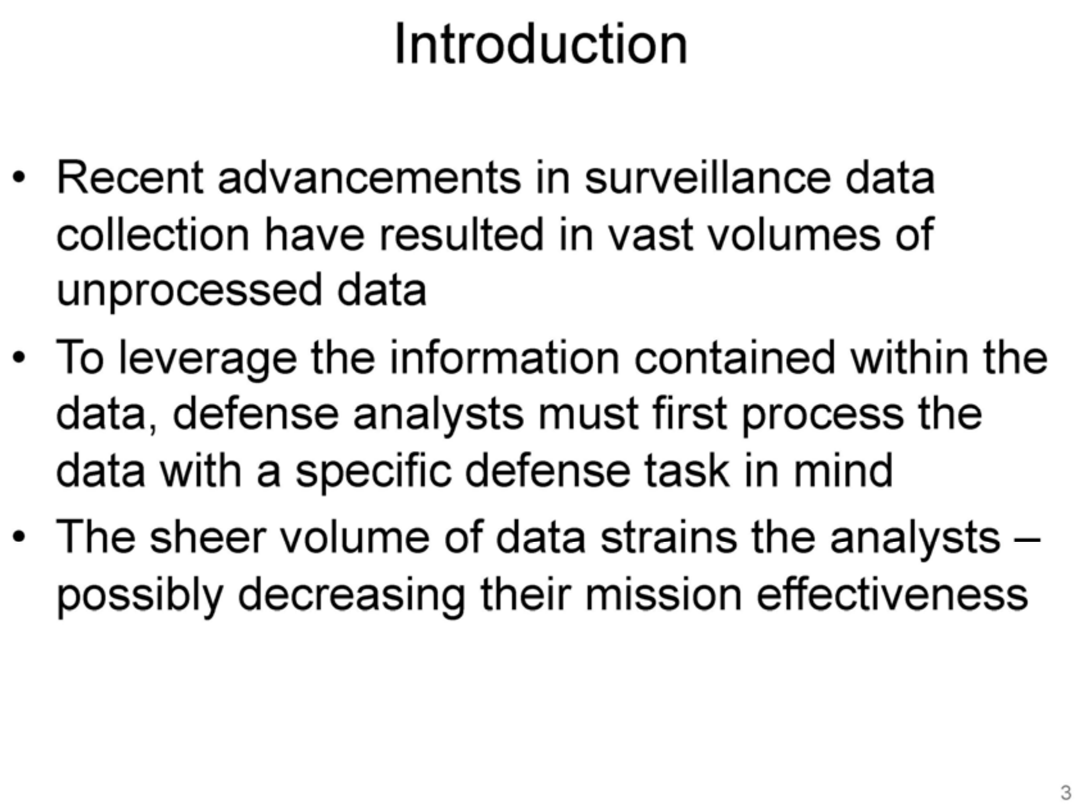
University of Virginia



Outline

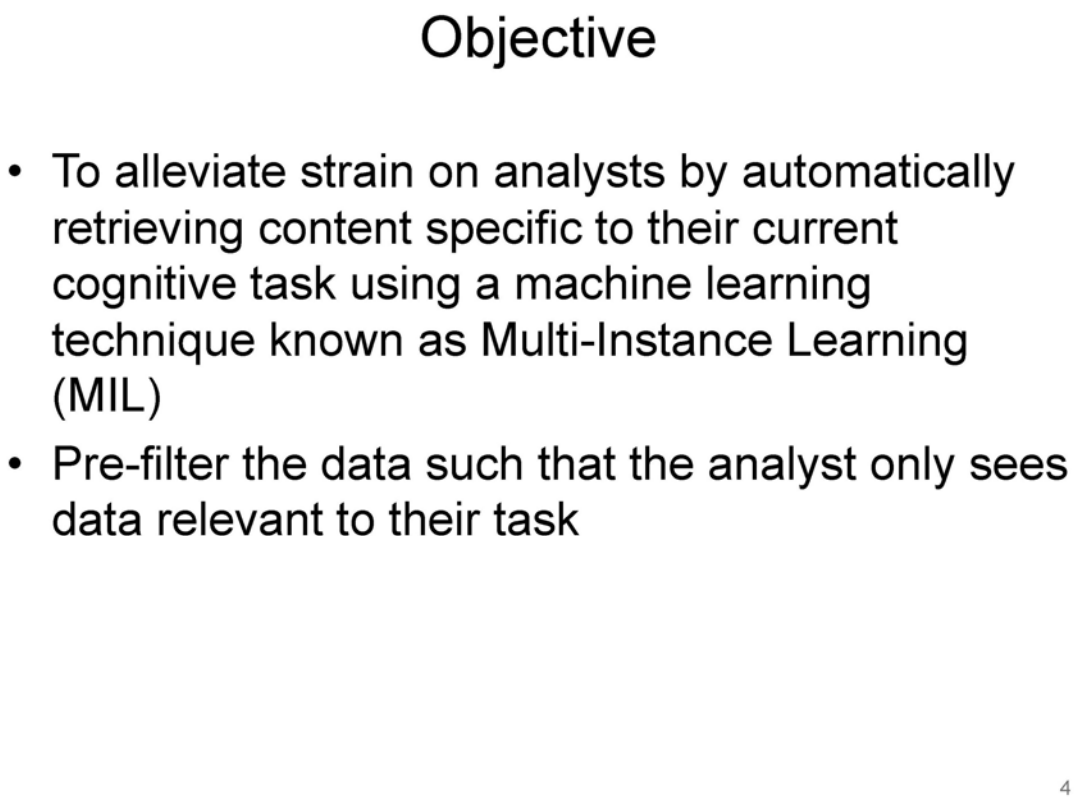
- Introduction
- Objective
- The Multi-Instance Learning (MIL) Algorithm
- Project Background
- Applying MIL to Vehicle Tracking
- Creating a Simulation
- Evaluating MIL
- Results
- Final Observations

2



Introduction

- Recent advancements in surveillance data collection have resulted in vast volumes of unprocessed data
- To leverage the information contained within the data, defense analysts must first process the data with a specific defense task in mind
- The sheer volume of data strains the analysts – possibly decreasing their mission effectiveness

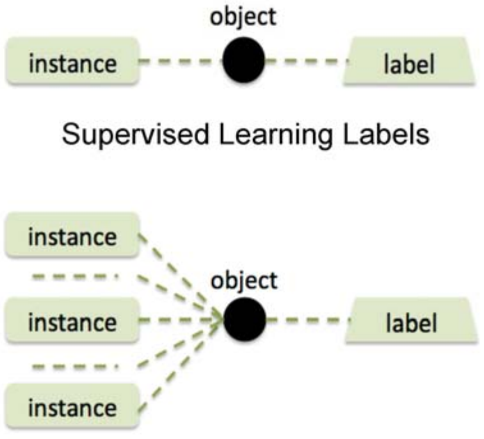


Objective

- To alleviate strain on analysts by automatically retrieving content specific to their current cognitive task using a machine learning technique known as Multi-Instance Learning (MIL)
- Pre-filter the data such that the analyst only sees data relevant to their task

About MIL

- A variation of supervised learning with ambiguity in **labels**
- A **bag** (object) can be represented in multiple ways – know as **instances**
- Each **bag** receives a **positive** or **negative label**, but it is unknown which **instance(s)** triggered the **label**.
- A bag will be labeled
  - positive if at least one instance in the bag is positive
  - negative if all instances in the bag are negative

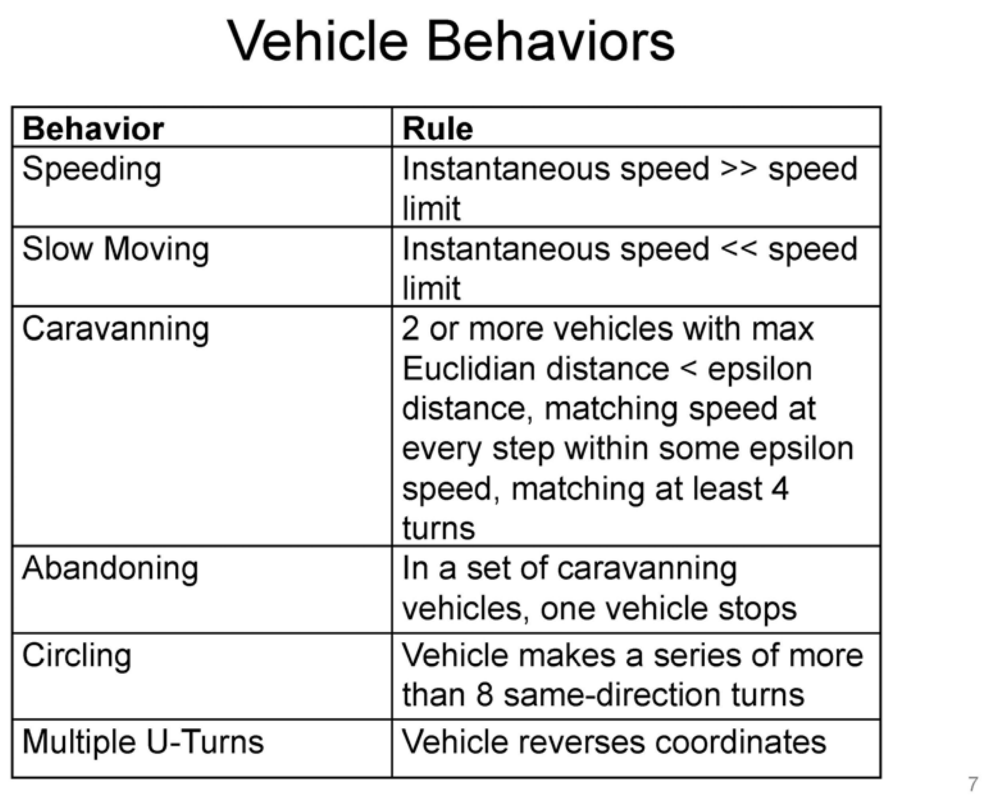


Supervised Learning Labels

Multi-Instance Learning Labels

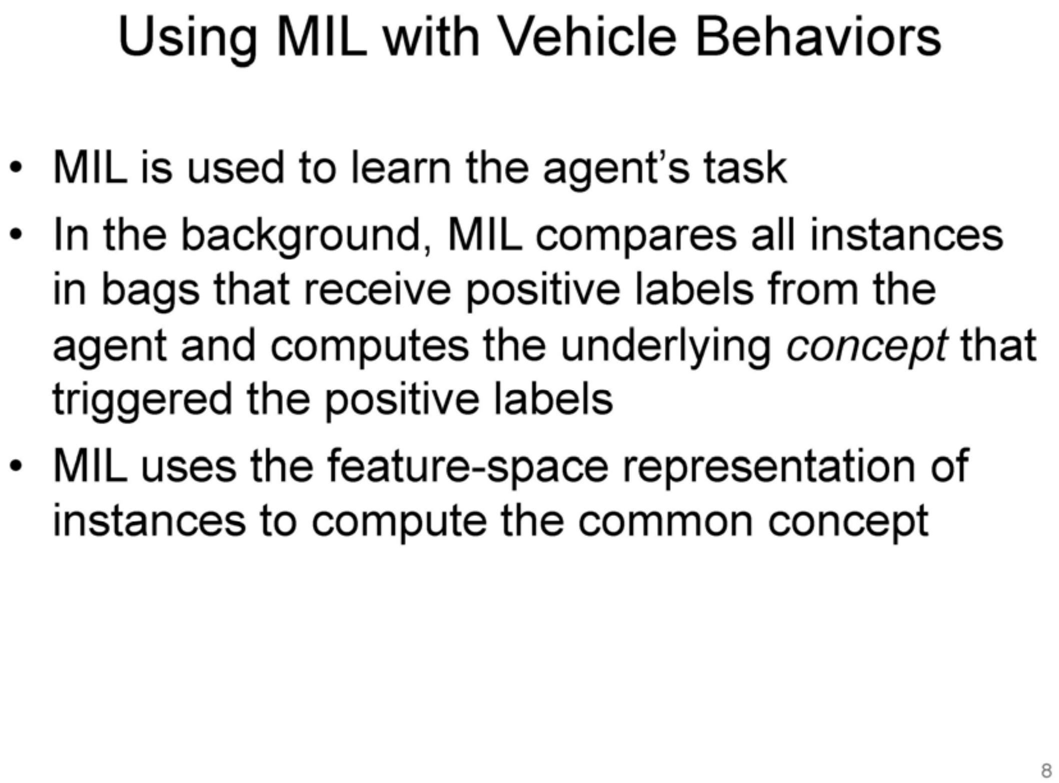
Some Project Background

- Picked vehicle tracking environment as proof-of-concept scenario
- Simulated environment models vehicle tracking activities
- Analyst is replaced by an agent
- Assigned the agent tasks that might be performed by an analyst tracking vehicles based on their behavior



Vehicle Behaviors

Behavior	Rule
Speeding	Instantaneous speed >> speed limit
Slow Moving	Instantaneous speed << speed limit
Caravanning	2 or more vehicles with max Euclidian distance < epsilon distance, matching speed at every step within some epsilon speed, matching at least 4 turns
Abandoning	In a set of caravanning vehicles, one vehicle stops
Circling	Vehicle makes a series of more than 8 same-direction turns
Multiple U-Turns	Vehicle reverses coordinates



Using MIL with Vehicle Behaviors

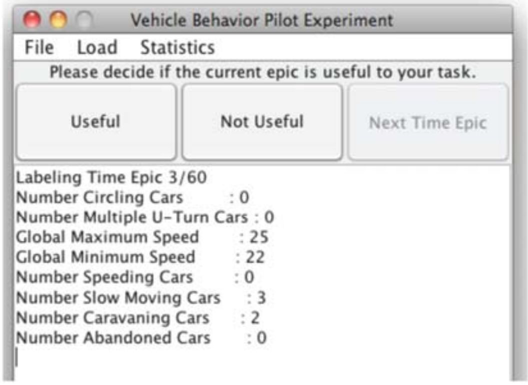
- MIL is used to learn the agent's task
- In the background, MIL compares all instances in bags that receive positive labels from the agent and computes the underlying *concept* that triggered the positive labels
- MIL uses the feature-space representation of instances to compute the common concept

Feature Space

Feature Name	Type
Speeding Cars Present	Boolean
Number Speeding Cars	Integer
Slow Moving Cars Present	Boolean
Number Slow Moving Cars	Integer
Caravans Present	Boolean
Number Caravans	Integer
Abandoned Cars Present	Boolean
Number Abandoned Cars	Integer
Circling Cars Present	Boolean
Number Circling Cars	Integer
Multiple U-Turn Cars Present	Boolean
Number Multiple U-Turn Cars	Integer
Number Cars	Integer
Number Trucks	Integer
Number Red Vehicles	Integer
Number Blue Vehicles	Integer
Global Maximum Speed	Integer
Global Minimum Speed	Integer

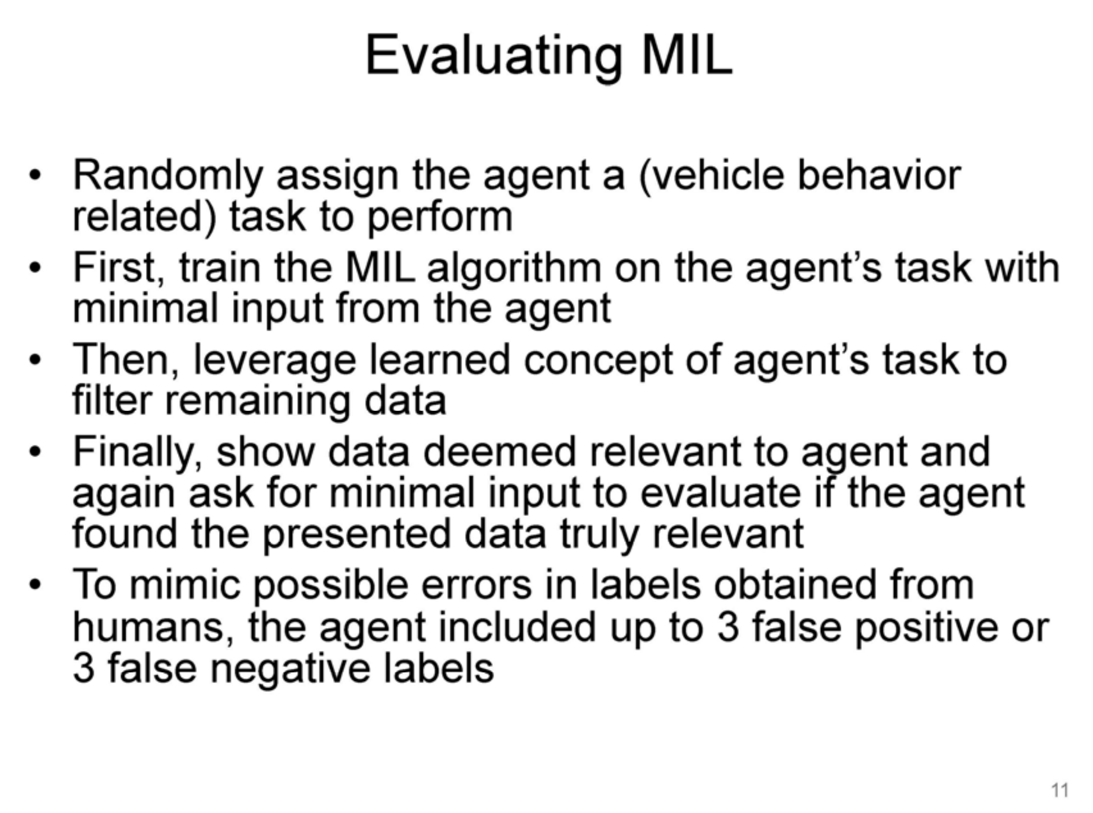
About the Simulation

- Used as a tool to evaluate the MIL algorithm
- Presents agent with randomly picked time epics of vehicle behavior
- Two main stages:
  - Training – MIL works in the background to learn the agent's task based on agent input
    - Agent input – answer to question: Was the information in the past time epic useful in fulfilling the task?
  - Evaluation – MIL filters data based on learned concept of agent's task
    - Simulation subsequently only presents relevant data unless performing an unaided trial



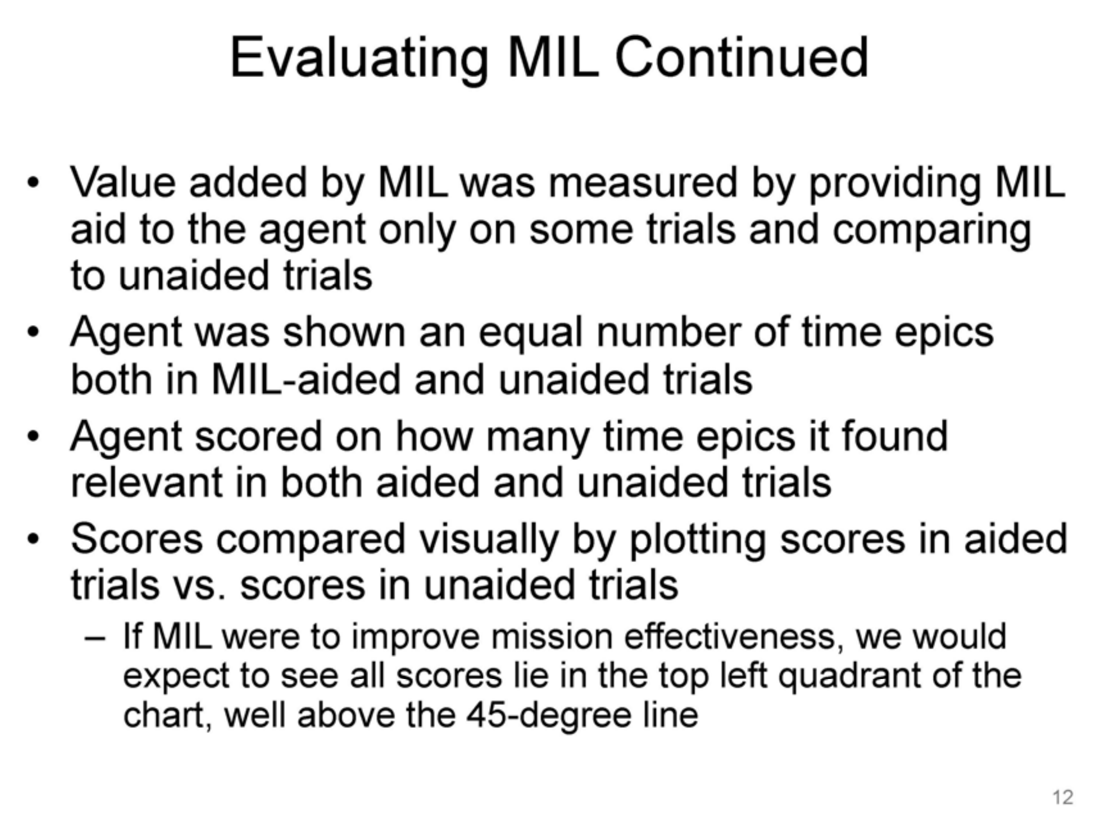
Please decide if the current epic is useful to your task.

```
Labeling Time Epic 3/60
Number Circling Cars        : 0
Number Multiple U-Turn Cars : 0
Global Maximum Speed        : 25
Global Minimum Speed        : 22
Number Speeding Cars        : 0
Number Slow Moving Cars     : 3
Number Caravaning Cars      : 2
Number Abandoned Cars       : 0
```



Evaluating MIL

- Randomly assign the agent a (vehicle behavior related) task to perform
- First, train the MIL algorithm on the agent’s task with minimal input from the agent
- Then, leverage learned concept of agent’s task to filter remaining data
- Finally, show data deemed relevant to agent and again ask for minimal input to evaluate if the agent found the presented data truly relevant
- To mimic possible errors in labels obtained from humans, the agent included up to 3 false positive or 3 false negative labels



Evaluating MIL Continued

- Value added by MIL was measured by providing MIL aid to the agent only on some trials and comparing to unaided trials
- Agent was shown an equal number of time epics both in MIL-aided and unaided trials
- Agent scored on how many time epics it found relevant in both aided and unaided trials
- Scores compared visually by plotting scores in aided trials vs. scores in unaided trials
  - If MIL were to improve mission effectiveness, we would expect to see all scores lie in the top left quadrant of the chart, well above the 45-degree line

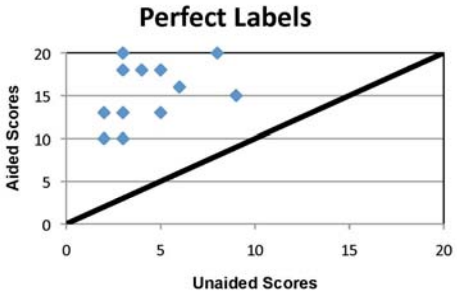
Agent Tasks

Task ID	Task Type	Task
1	Simple	Identify ABANDONED vehicles
2	Simple	Identify CARAVANS
3	Simple	Identify CIRCLING vehicles
4	Simple	Identify SLOW vehicles
5	Simple	Identify SPEEDING vehicles
6	Simple	Identify U-TURN vehicles
7	Composite 1/2	Identify ABANDONED & CARAVANS
8	Composite 1/2	Identify CIRCLING & U-TURNS
9	Composite 1/2	Identify SLOW & SPEEDING
10	Composite 1/2	Identify SPEEDING & CARAVANS
11	Composite 2/3	Identify CARAVANS & ABANDONED & SLOW
12	Composite 2/3	Identify SLOW & CIRCLING & U-TURNS
13	Composite 2/3	Identify SLOW & SPEEDING & U-TURNS
14	Composite 2/3	Identify SPEEDING & CARAVANS & U-TURNS

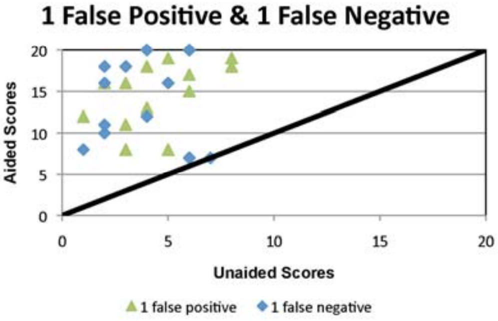
Results

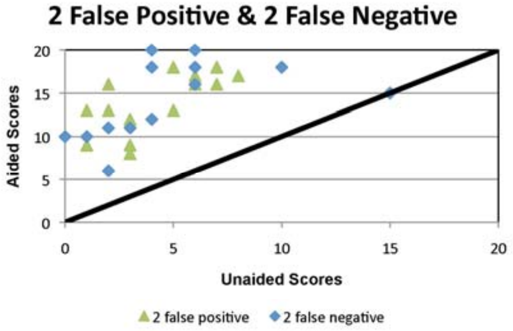
- The agent received higher scores with the aid of MIL!
- Lift shown here implied improved mission effectiveness

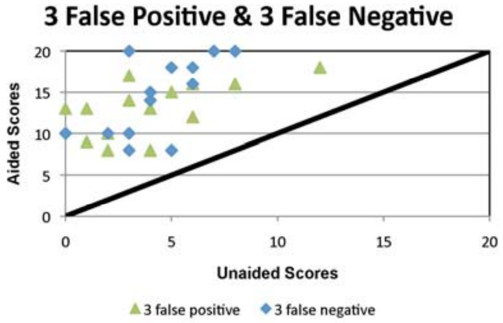
	Score	
Label Type	Aided	Unaided
Perfect	15.50	4.57
1 False Positive	14.14	4.35
1 False Negative	13.79	3.71
2 False Positive	13.93	4.21
2 False Negative	14.64	4.79
3 False Positive	13.00	4.07
3 False Negative	14.07	4.14



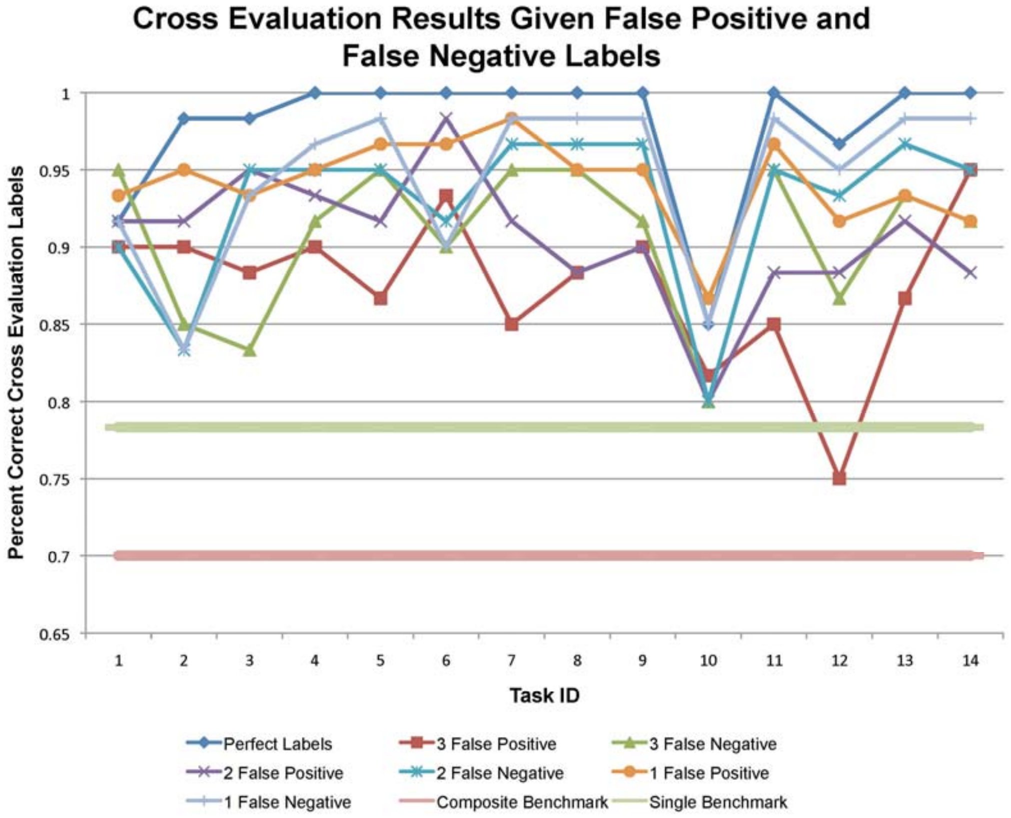
Results – Type I & Type II Errors


1 False Positive & 1 False Negative

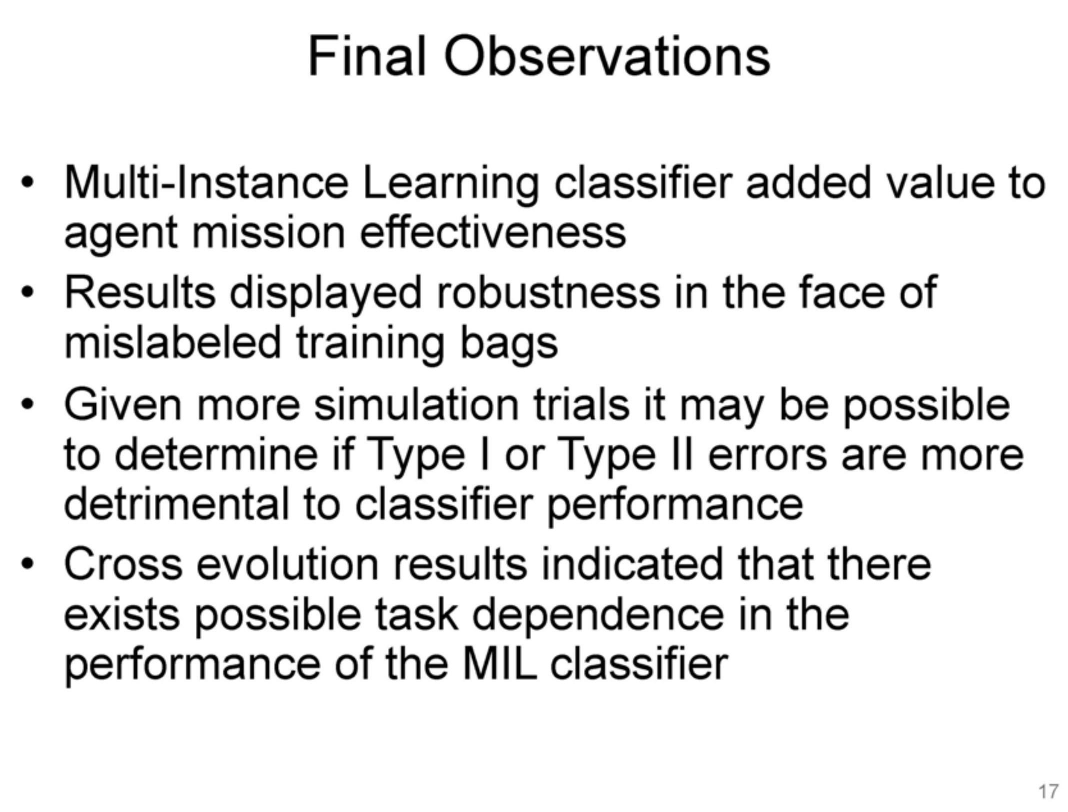

2 False Positive & 2 False Negative


3 False Positive & 3 False Negative

15


Cross Evaluation Results Given False Positive and False Negative Labels

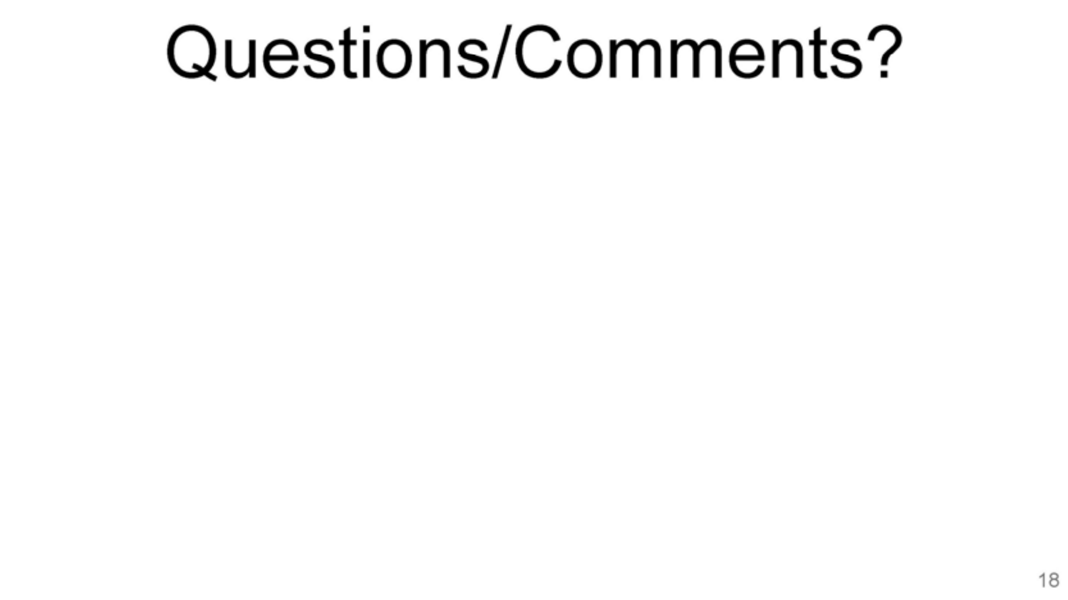
16



Final Observations

- Multi-Instance Learning classifier added value to agent mission effectiveness
- Results displayed robustness in the face of mislabeled training bags
- Given more simulation trials it may be possible to determine if Type I or Type II errors are more detrimental to classifier performance
- Cross evolution results indicated that there exists possible task dependence in the performance of the MIL classifier

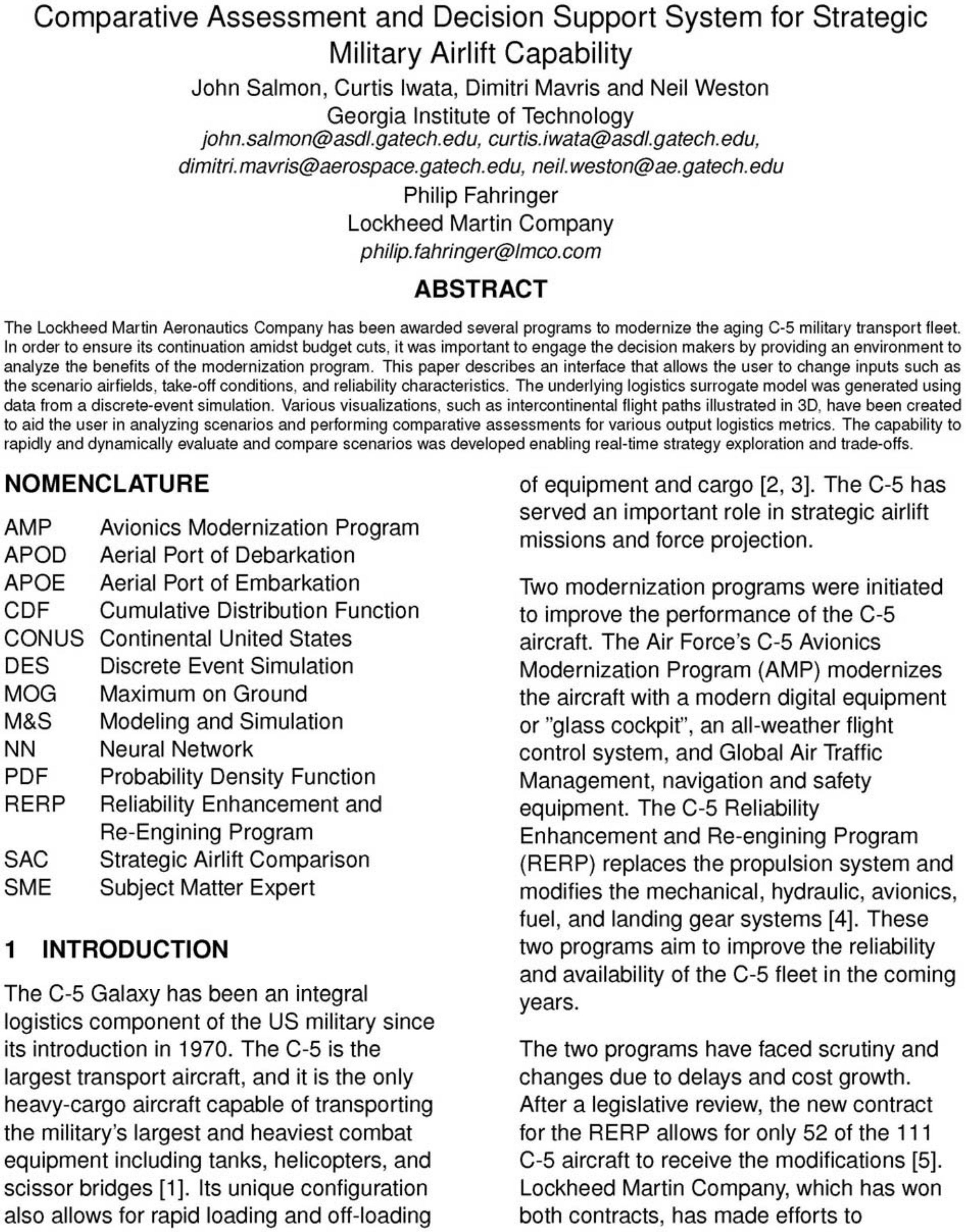
17



Questions/Comments?

18

1.6 Comparative Assessment and Decision Support System for Strategic Military Airlift Capability



Comparative Assessment and Decision Support System for Strategic Military Airlift Capability

John Salmon, Curtis Iwata, Dimitri Mavris and Neil Weston
Georgia Institute of Technology
john.salmon@asdl.gatech.edu, curtis.iwata@asdl.gatech.edu,
dimitri.mavris@aerospace.gatech.edu, neil.weston@ae.gatech.edu
Philip Fahringer
Lockheed Martin Company
philip.fahringer@lmco.com

ABSTRACT

The Lockheed Martin Aeronautics Company has been awarded several programs to modernize the aging C-5 military transport fleet. In order to ensure its continuation amidst budget cuts, it was important to engage the decision makers by providing an environment to analyze the benefits of the modernization program. This paper describes an interface that allows the user to change inputs such as the scenario airfields, take-off conditions, and reliability characteristics. The underlying logistics surrogate model was generated using data from a discrete-event simulation. Various visualizations, such as intercontinental flight paths illustrated in 3D, have been created to aid the user in analyzing scenarios and performing comparative assessments for various output logistics metrics. The capability to rapidly and dynamically evaluate and compare scenarios was developed enabling real-time strategy exploration and trade-offs.

NOMENCLATURE

AMP	Avionics Modernization Program
APOD	Aerial Port of Debarkation
APOE	Aerial Port of Embarkation
CDF	Cumulative Distribution Function
CONUS	Continental United States
DES	Discrete Event Simulation
MOG	Maximum on Ground
M&S	Modeling and Simulation
NN	Neural Network
PDF	Probability Density Function
RERP	Reliability Enhancement and Re-Engining Program
SAC	Strategic Airlift Comparison
SME	Subject Matter Expert

1 INTRODUCTION

The C-5 Galaxy has been an integral logistics component of the US military since its introduction in 1970. The C-5 is the largest transport aircraft, and it is the only heavy-cargo aircraft capable of transporting the military's largest and heaviest combat equipment including tanks, helicopters, and scissor bridges [1]. Its unique configuration also allows for rapid loading and off-loading of equipment and cargo [2, 3]. The C-5 has served an important role in strategic airlift missions and force projection.

Two modernization programs were initiated to improve the performance of the C-5 aircraft. The Air Force's C-5 Avionics Modernization Program (AMP) modernizes the aircraft with a modern digital equipment or ”glass cockpit”, an all-weather flight control system, and Global Air Traffic Management, navigation and safety equipment. The C-5 Reliability Enhancement and Re-engining Program (RERP) replaces the propulsion system and modifies the mechanical, hydraulic, avionics, fuel, and landing gear systems [4]. These two programs aim to improve the reliability and availability of the C-5 fleet in the coming years.

The two programs have faced scrutiny and changes due to delays and cost growth. After a legislative review, the new contract for the RERP allows for only 52 of the 111 C-5 aircraft to receive the modifications [5]. Lockheed Martin Company, which has won both contracts, has made efforts to emphasize the need to keep the programs in place and to promote the need to upgrade the remaining fleet.

One such effort described in this paper is the development of the Strategic Airlift Comparison (SAC) tool. The SAC tool shows the logistical advantage of having a C-5M fleet over a corresponding C-5A fleet. The tool leverages modeling and simulation (M&S) to present a convincing argument in support of the modifications in a format that can be easily understood by decision makers such as government officials and military officers who may need additional support for their decision making process.

The paper begins with the development of the M&S environment to support the evaluation and decision making process of the C-5 modernization programs. A discrete-event simulation (DES) was created to generate the data for metrics including time to close a specific mission scenario and fuel consumption. The second half of the paper describes the interface which allows the user to display and manipulate the data in order to make comparisons between the different platforms. Different types of visualizations are also presented and discussed. The paper concludes with several examples of how this tool can be used to make comparisons.

2 DECISION MAKING AND M&S

Trends in cost overruns in defense acquisition have put accountability as a priority. Acquisition decisions must be traceable and should be backed by sound technical analysis. Since, decision makers often require insight into the technical details of the problem, they typically consult subject matter experts (SME). However, even when the SMEs are included in the discussions, there are situations when the experience of the SMEs is not adequate to answer speculative questions and ”what if scenarios”. For these questions, M&S can provide valuable insight into the problem at affordable costs and within a limited time frame [6].

2.1 Discrete Event Simulation

A discrete-event simulation (DES) was created to simulate the performance of a C-5 fleet. The DES models a system as a series of events, and this is well-suited for a logistics problem. For example, an aircraft arriving at an airport is an event, and each of these actions can be modeled per aircraft. Then a group of aircraft can be simulated together to gain insight on how an entire fleet would perform.

Several DES programs were examined including ExtendSim, Process Simulator, SimEvents and SimPy. ExtendSim and Process Simulator are commercial DES packages, and the users graphically arrange the process elements and the linkages [7, 8]. Process Simulator also requires Microsoft Visio to run. SimEvents is a DES package for MATLAB and also uses a graphical interface [9]. Simulation in Python (SimPy) is an object-oriented DES package written in the Python programming language [10]. It is open source and executes quickly; however, the set up and the outputs of the simulation are not graphical. For this project, SimPy was selected for its flexibility, fast run time, and availability.

The selection of SimPy proved to be prudent when the simulation runs had to be scaled up and distributed over several workstations. The open source nature allowed the team to run cases on different machines without worrying about licensing issues. Furthermore, multithreading was enabled using the Parallel Python package, allowing the program to use all processor cores.

2.2 Simluation Set Up

The simulation captures a simple logistics scenario of transporting cargo from the

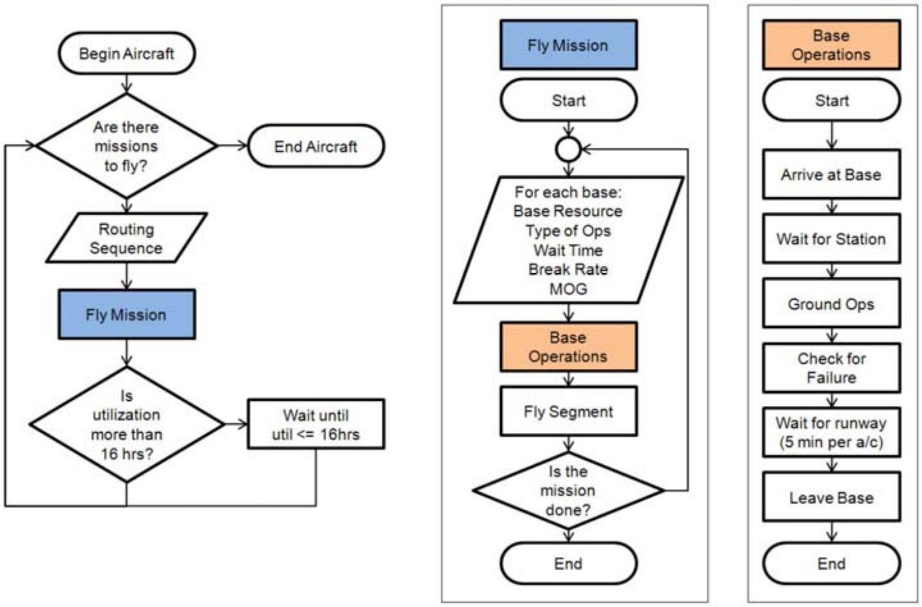


Figure 1. Simulation Flow Block Diagram

aerial port of embarkation (APOE) to the aerial port of disembarkation (APOD). The simulation flow is outlined in Fig. 1.

There are four different routing options available which are the following: 1) direct flight between APOE and APOD, 2) en route location for forward flight with direct return flight, 3) same en route location for forward and return flights, and, 4) two different en route location for forward and return flights.

The simulation begins with all of the aircraft at the APOE, and it ends when the last aircraft returns back to the APOE. At each airfield, a maximum on ground (MOG) constraint limits the number of aircraft that can be serviced at any given time. Thus, if an airfield has a MOG of three and has five aircraft on site, two of the aircraft would have to wait on the side of the tarmac until a station opens up.

In the service station, the aircraft waits for a set amount of time for the scheduled activity at the airfield, such as loading cargo and refueling. Time is dependent on the type of aircraft and activity. The ground times listed in the AFPAM10-1403 were used [3].

After the servicing is completed, a random probabilistic algorithm is used to determine if the aircraft had suffered any minor failures during its previous flight and how long it would take to repair it. All failures are assumed to take no more than 72 hours, and the repair time of each failure is determined based off of a pre-defined random distribution. The aircraft then prepares for take-off, and only one aircraft can occupy the runway. If there are multiple aircraft ready to take-off, they must wait for their turn.

The flight times are calculated based on the aircraft's block speed. Within the simulation, the same type of aircraft will fly the equivalent distance in the same amount of time. The distances have been calculated using the great circle distance formula, and the variability of flight times due to weather and other factors were not incorporated into this simulation.

Utilization is a metric of how hard the crew works, and this is typically calculated as an average amount of hours flown per day. For the simulation, the utilization was calculated as the ratio of the total flight time, from leaving the APOE to returning back to the APOE, and the total amount of time spent since leaving the APOE. A cap of 16 hours was placed, and if any aircraft exceeded this at the end of its mission, it was grounded before returning to service at the APOE.

2.3 Model Inupts

The DES does not require the physics of the problem, such as the amount of cargo and the amount of fuel consumed. Much of this information was calculated beforehand, and the model was abstracted out, reducing the number of inputs and increasing its extensibility to other aircraft. The input variables into the DES are listed in Table 1.

Table 1. SimPy Model Input Variables

SimPy Inputs	Min	Max
Type of Aircraft	0	1
Fleet Size	1	70
Number of Flights	1	1000
Routing	1	4
Flight Time: Leg 1	30	1200
Flight Time: Leg 2	30	1200
Flight Time: Leg 3	30	1200
Flight Time: Leg 4	30	1200
Repair Prob: 0 to 4 hrs	0	1
Repair Prob: 4 to 12 hrs	0	1
Repair Prob: 12 to 24 hrs	0	1
Break Rate	0.01	0.6
MOG: APOE	1	15
MOG: APOD	1	15
MOG: En Route 1	1	15
MOG: En Route 2	1	15

2.4 Model Limitations

The DES model currently supports the C-5 and the C-17. Adding other types of aircraft is possible as long as the ground servicing times are available. However, this only pertains to running the model itself, and the payload-range curve, block speed and fuel burn are needed to be useable with the interface.

Another current limitation is that the model has only four options for routes. In reality, there may be more airfields involved in each mission. For example, refueling usually does not occur at the APOD if it is in a hostile location; a recovery airfield close to and safer than the APOD would be involved. Including more airfields makes the routing more complicated and increases the number of variables.

There have also been discussions about extending capability to hybrid fleets where different types of aircraft would fly different legs of the mission. For example, the C-5 could fly transatlantic flights, while the C-17 could fly the shorter legs to different airfields in the region.

2.5 Simulation Runs

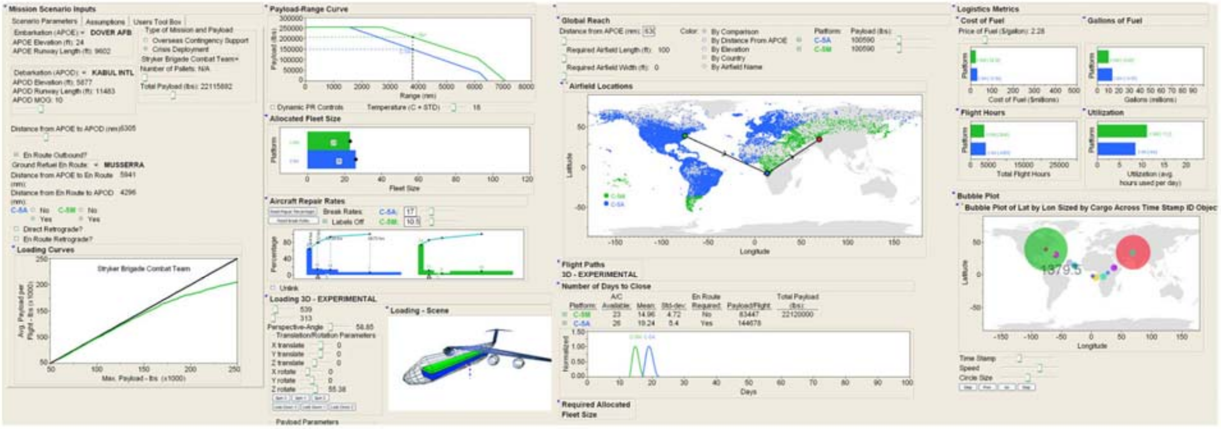
Over 50,000 cases were used to populate the design space. A combination of central composite and space filling designs were used to generate the cases to run. Each case was repeated 1000 times in order to generate the distribution due to the variability of failure events and repair times. The runs were allocated across several workstations.

Each case of a 1000 runs took anywhere from a few seconds to over 30 seconds depending on the input parameters, and the final set of data took a few days to run across several machines.

2.6 Surrogate Modeling

Surrogate models of the DES model were created to avoid carrying the large data set and to enable interactivity with the of the interface.

The design space was nonlinear, and the time to close increased dramatically as the fleet size decreased to one aircraft. A neural network (NN) model was fitted to the data, but the NN model outputs at the edges had poor fits due to the nonlinearity of the data. To resolve this issue, a separate NN was



Embarkation (APOE): DOVER AFB
APOE Elevation (ft): 24
APOE Runway Length (ft): 9602

Debarkation (APOD): KABUL INTL
APOD Elevation (ft): 5877
APOD Runway Length (ft): 11483
APOD MOG: 10

○ Overseas Contingency Support
● Crisis Deployment
Stryker Brigade Combat Team
Number of Pallets: N/A
Total Payload (lbs): 22115882

Distance from APOE to APOD (nm):6305

☒ En Route Outbound?
Ground Refuel En Route: MUSSERRA
Distance from APOE to En Route (nm): 5941
Distance from En Route to APOD (nm): 4296
C-5A ○ No ● Yes
C-5M ○ No ● Yes
☐ Direct Retrograde?
☐ En Route Retrograde?

☐ Dynamic PR Controls
Temperature (C + STD): 18

Break Rates: C-5A: 17 C-5M: 10.5
☒ Labels Off
☐ Unlink

539
313
Perspective-Angle 58.85
Translation/Rotation Parameters
X translate 0
Y translate 0
Z translate 0
X rotate 0
Y rotate 0
Z rotate 55.38

Distance from APOE (nm): 63
Required Airfield Length (ft): 100
Required Airfield Width (ft): 0

Color: ● By Comparison
○ By Distance From APOE
○ By Elevation
○ By Country
○ By Airfield Name

Platform:	Payload (lbs):
C-5A	100590
C-5M	100590

Platform:	A/C Available:	Mean:	Std-dev:	En Route Required:	Payload/Flight:	Total Payload (lbs):
C-5M	23	14.96	4.72	No	83447	22120000
C-5A	26	19.24	5.4	Yes	144678	

Price of Fuel ($/gallon): 2.28

Figure 2. Strategic Airlift Comparison Tool Layout

fitted to the data for low number of flights and small fleet sizes. The two NN models were patched together using linear interpolation.

3 DECISION SUPPORT INTERFACE

3.1 Interface Tool Layout

The decision support interface is developed within JMP, one of the statistical software packages produced by the SAS institute [11]. The layout of the tool is arranged with inputs at the left or top and outputs to the right or at the bottom. Inputs are entered or selected in four outline boxes: Mission Scenario Inputs, Payload-Rage Curve, Allocated Fleet Size, and Aircraft Repair Rates. Output outline boxes include Global Reach with Airfield Location, Number of Days to Close, Required Allocated Fleet Size and Logistics Metrics. The layout is adjustable for various screen sizes and resolutions and is shown for a wide screen format in Fig. 2. Some of these various outline boxes are described briefly in the following sections. In addition, a number of the outline boxes contain advanced visualizations that assist the user in understanding specific scenarios, or comparing between aircraft platforms.

3.2 Mission Scenario Inputs

Application and analysis of the tool will typically begin within the Mission Scenario Input outline box at the top right in Fig. 2. Mission scenarios are defined by three essential inputs, namely, the APOE, the APOD and a given mission payload. The other input parameters are defaulted to initial values, but can be adjusted. For example, the type of payload is adjustable to five different payload types typically transported on C-5 aircraft. When selecting a payload type (other than the default custom payload option), the loading curve will adjust accordingly and will represent the actual average payload per flight based on constraints such as temperature, airfield length, cubing-out conditions, etc.

Since some missions will require an en route location for refueling, or for increasing average payloads per mission, the various options for adding such an airfield are placed below the APOE and APOD selection boxes. Three options for en route locations include: Outbound, En Route Retrograde, and Direct Retrograde and is provided for the respective situations of refueling on the way to the APOD, on the way back to the APOE, or flying direct back to the APOE without refueling. In addition to selecting the en route locations, comparisons between the C-5A and C-5M platforms are possible by having one or both stop over in different combinations. This allows the user to quickly perform what-if comparisons and investigate the advantages and disadvantages of stopping to refuel at different locations, for different distances, among the various constraints.

The two other tabs within this outline box, (i.e. Assumptions and Users Tool Box) allow access to additional settings and constraints for the airfields and visualizations respectively. Within the Assumptions tab, users can modify the MOG for each airfield and within the Users Tools Box, the layout can be adjusted, colors can be changed and example Bubble Plots can be created which are described later.

3.3 Payload-Range Curve, Fleet Size and Repair Rates

The payload-range (PR) curve displayed corresponds with the selected APOE, by making use of the elevation and airfield length data at that particular location. Furthermore, in calculating the PR curve the ambient temperature at the APOE can be set from 5 to 40 °C making the curve dynamic for this temperature range. Other settings can account for variable reserve fuels and minimum payload per flight requirements set by toggling the Dynamic PR Controls check box.

The fleet size for each of the C-5 platforms is entered using an interactive bar chart within the Allocated Fleet Size outline box. The values themselves are both linked and constrained in that there are only 59 C-5As currently available, and only 111 C-5Ms possible for allocation if all 59 C-5As, and the other 52 C-5B and Cs were modified. This useful technique keeps scenarios and comparisons as defined by the user within the realm of possibility.

The Aircraft Repair Rates outline box allows the user to adjust reliability parameters for the aircraft. First, break rates, defined as the percentage of flights that require repair (and therefore a delay) for each leg of the mission are set for each platform in a range from 0 to 60%. Next the probability for each of the four categories of repair or delay times are set. The percentage of time that repairs will fall into one of the four delay time categories, (i.e. 0-4hrs, 4-12hrs, 12-24hrs, and 24-72hrs) are positioned by the user dragging the marker for each category's bar, or by adjusting the cumulative distribution function, which are internally linked.

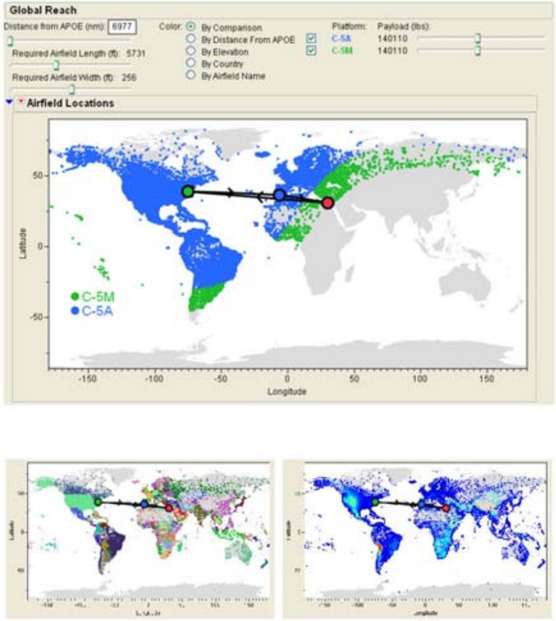
The average repair time for the given distribution is indicated at the bottom and is used to generally compare a platform's recoverability. For example, a platform's lower average repair time, compared to the other will be more recoverable, characterized by low repair times, even with a higher break rate. The overall reliability is thus a function of the repair times for each category and the break rate which are used within and surrogate model of the the discrete-event simulation.

3.4 Number of Days to Close and Logistic Metric Outputs

Although there are two types of missions selectable within the scenario inputs outline box, namely, "Overseas Contingency Support" and "Crisis Deployment", the user will be more often interested in the latter. Thus, the output metric of interest is the number of days to close a mission or the time required to deliver the specified total mission payload to the APOD. The mean, standard deviation and shape of the time to close distribution is presented in the table and graph within the Number of Days to Close outline box. These output metrics are updated in real-time by changing any of the other input parameters discussed previously. Any modification to the scenario, fleet size, or reliability parameters will automatically rerun a surrogate model to recreate the time to close distribution and other output logistic metrics.

Other logistic metrics, including cost and gallons of fuel, flight hours, and utilization are traded against the time to close output metric. Clearly, closing the mission as soon as possible is desirable, but the cost of fuel and flight hours (i.e. a representation of human resources required) may be too. Similarly, the total mission payload could be delivered in a very short amount of time by increasing the fleet size, but this too may be impractical or even impossible. Thus, the trade-offs and sensitivities between output metrics, but also between inputs and outputs, becomes examinable.

3.5 Airfield Selection and Filtering



Global Reach

Distance from APOE (nm): 6977

Required Airfield Length (ft): 5731

Required Airfield Width (ft): 256

Color: ● By Comparison
○ By Distance From APOE
○ By Elevation
○ By Country
○ By Airfield Name

	Platform:	Payload (lbs):
☑	C-5A	140110
☑	C-5M	140110

Airfield Locations

Figure 3. Examples of Airfield Filtering and Coloring Schemes

The visualization of the user-defined scenario is automatically displayed on the airfield locations map. However, before the APOE and other scenario airfields are selected, the map also allows the user to filter, down select and investigate airfield locations and attributes. Two filtering methods are provided as defaults: the required airfield length and required airfield width, and are set to remove or hide all airfields that do not meet those requirements.

Further investigation can be performed by right clicking on any airfield location and exploring the airfield attributes. Other useful coloring schemes, such as coloring by elevation or country, are available, along with the JMP Data Filter which can be implemented to add constraints or other filtering options (See Fig. 3). The coloring "By Distance From APOE" and "By Comparison" is only active after selecting an APOE. This latter option allows comparisons between the different platforms for reachable airfields from the selected APOE, and can quickly illustrate the capability gap of the C-5Ms over the C-5As for equal payloads and take-off conditions.

3.6 Advanced Visualizations and Analyses

A number of visualizations were developed to answer concerns identified during various test phases of the tool's development. These visualizations aid decision makers in understanding the assumptions and details of the scenario, and clarify certain model characteristics and outputs.

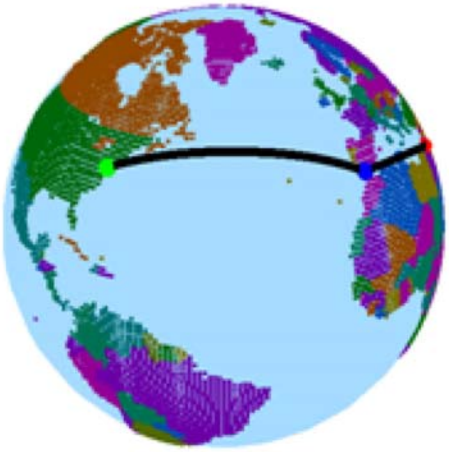


Figure 4. Visualization of 3D Flight Paths

One of most commonly expressed limitations, for example, was the potential for users to be confused with the flight path representation on the 2D map. Developers of the tool decided that a simple node and edge diagram for the flight paths (with only straight edges) would be most clear in a graphical representation of the mission scenario. The disadvantage of such a decision is the possible confusion of the actual flight path taken by aircraft since the true path on a flat earth model will be curved along the great circle distance between two points. To avoid this possibility, the actual path taken between two airfields is redrawn onto a 3D earth model as shown in Fig. 4.

The advantages of this visualization include mission leg distances that are relatively consistent with each other (not always the case for projections onto a 2D map) and the rapid identification of airspace violations against countries or other political constraints. Enhancements to this visualization will include analysis of how the flight path will circle around a particular nation, which will not allow C-5 aircraft from entering its airspace and the resultant impact on the output and logistic metrics.

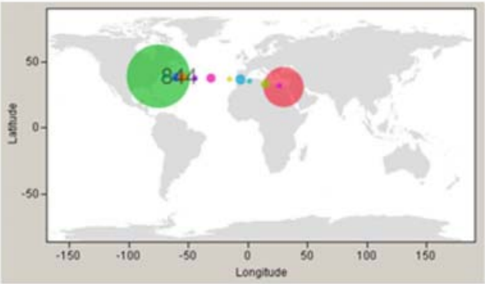


Figure 5. Bubble Plot Animation Showing the Movement of Payload

To assist decision makers in understanding the delivery of the mission payload over time and how aircraft are delayed and interact with each other at the various airfields, a Bubble Plot representation of the mission can be executed from the Users Tool Box tab in the Mission Scenario Inputs outline box (See Fig. 5).

The location of all aircraft and the current location of the payload (i.e. at the APOE, in flight or at the APOD), with relative time stamps can be explored at all times during the mission. Furthermore, comparisons in time of different scenarios, fleet sizes, etc. can be performed with more than one Bubble Plot. To enable this comparison, the Bubble Plot is simply executed again after a change in the mission scenario (such as a new en route location), and then setting the speed and time stamp values within the plots consistently. Exploring the delivered payload over time can be helpful in assisting decision makers to identify the impact of parameter changes on time to close or other metrics when the scenario is changed at various times in the mission.

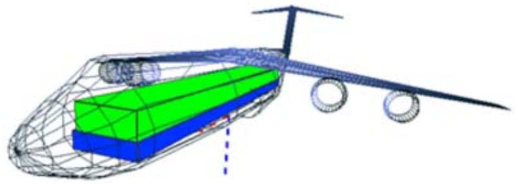


Figure 6. 3D Loading Volume Representation (blue: C-5A, green: C-5M)

Another visualization, currently in the experimental phase of development, is a 3D representation of the loading volume of the two platforms (See Fig. 6). When adjusting the temperature of the PR curve or changing the location of the APOE, the maximum possible payload weight will fluctuate in order to satisfy the constraint of the mission scenario and the platform itself. To compare the cargo transporting holding capabilities of the C-5M over the C-5A, a volume representing the percentage of the maximum payload weight is projected within the cargo bay of the aircraft. These volumes represent the weight percentage and currently not the actual space used by any particular payload type. Potential enhancements to this visualization include detailed cargo characterization and how loading schemes affect the logistic metrics and other performance and aerodynamic metrics such as trim angles and stability points. Thus, future analyses may allow the decision maker to trade not only programmatic and mission parameters with high level logistic metrics, but potentially with lower level performance and operational metrics as well.

4 DISCUSSION

Using the input parameters and the DES surrogate model, trade-offs and analyses can become interactive and provide more insight. For example, suppose that a decision maker wants to compare the performance of the C-5A and C-5M for a particular cargo delivery mission. In this scenario, the C-5A will be making a stopover at an en route location, while the C-5M will fly directly to and from the APOD. In terms of time to close, the C-5A may complete the mission faster, but there is an additional cost for operating a refueling base overseas. For the decision maker, the cost savings of the C-5M may be the better option despite the longer time to close. In another scenario, there may be a hard requirement on time to close, and the fleet size can be changed to satisfy it. Furthermore, the two platforms will fly the same route so no overseas bases are needed. In this case, time is saved at the expense of maintaining a larger fleet at home.

The addition of changeable break rates and repair rates provides insight into how reliability affects the fleet size required to complete a mission. For many scenarios, reducing the break rates can decrease the average time to close and decrease the distribution. Likewise, fleet size and repair rates can be traded against the other logistic parameters. Utilization may be higher with more reliable aircraft, but lower with larger fleet sizes. Decision makers can also use the tool to investigate what targets for improved reliability are required to increase a particular capability. For example, what level of break rate must be reached to decrease the time to close by 10%?

Total cost of fuel can also be traded with a different en route location. An en route location with a higher MOG may be more “out of the way,” but it may be less risky for bottle necks if repairs are necessary. Thus, although the cost of fuel will be greater, the time to close will likely be lower by using the base with a higher MOG constraint.

Lastly, all of these outputs and parameters are dependent on the PR curve and on the underlying assumptions of required reserve fuel and temperature. At some temperatures, the difference in time to close for the C-5A and C-5M is significant and favors the re-engined C-5M, but at lower temperatures the capability gap isn’t quite as significant. Reserve fuel is also available for adjustment. This parameter (similar to the risk of low MOG airfields) allows the decision maker to apply different levels of risk to the scenario by exploring the impact of reducing the required reserve fuel on metrics such as time to close and fuel cost.

5 CONCLUSION

An interactive tool was developed to provide decision makers access to simulation models and data for improved analysis, mission planning and aircraft platform comparisons. Surrogate modeling was a key component for rapidly and dynamically performing trades between various logistic metrics including cost, risk and time. The surrogate model was created from discrete-event simulations of military cargo missions delivering payloads to airfields around the world, across a range of input attributes and parameters. Sensitivity analyses, what-if games and comparisons between aircraft platforms are enabled in real-time through a combination of dynamic visualizations, adjustable assumptions, and surrogate modeling techniques.

References

[1] USAF, “Factsheets: C-5 galaxy.” ”http://www.af.mil/information/

factsheets/factsheet.asp?id=84", June 2009.

[2] GlobalSecurity, "C-5 galaxy." "http://www.globalsecurity.org/military/systems/aircraft/c-5.htm", April 2006.

[3] USAF, "Air force pamphlet 10-1403: Air mobility planning factors," 2003.

[4] GAO, "Defense acquisitions. assessments of selected weapon programs," Tech. Rep. March, Government Accountability Office, 2009.

[5] C. Bolkcom and W. Knight, "Strategic airlift modernization: Analysis of C-5 modernization and C-17 acquisition issues," tech. rep., Congressional Research Service, 2008.

[6] M. Semini, H. Fauske, and J. O. Strandhagen, "Applications of discrete-event simulation to support manufacturing logistics decision-making: a survey," in *Proceedings of the 38th conference on Winter simulation*, WSC '06, pp. 1946–1953, Winter Simulation Conference, 2006.

[7] ExtendSim, "Extendsim simulation software by imagine that inc.." "http://www.extendsim.com/", June 2010.

[8] ProModel, "Process simulator." "http://www.promodel.com/products/processsimulator/", June 2010.

[9] MathWorks, "Simevents - discrete event simulation software - simulink." "http://www.mathworks.com/products/simevents/", June 2010.

[10] SimPy, "Simpy simulation package homepage." "http://simpy.sourceforge.net/", June 2010.

[11] SAS, "JMP (John's Macintosh Program)," 2008.

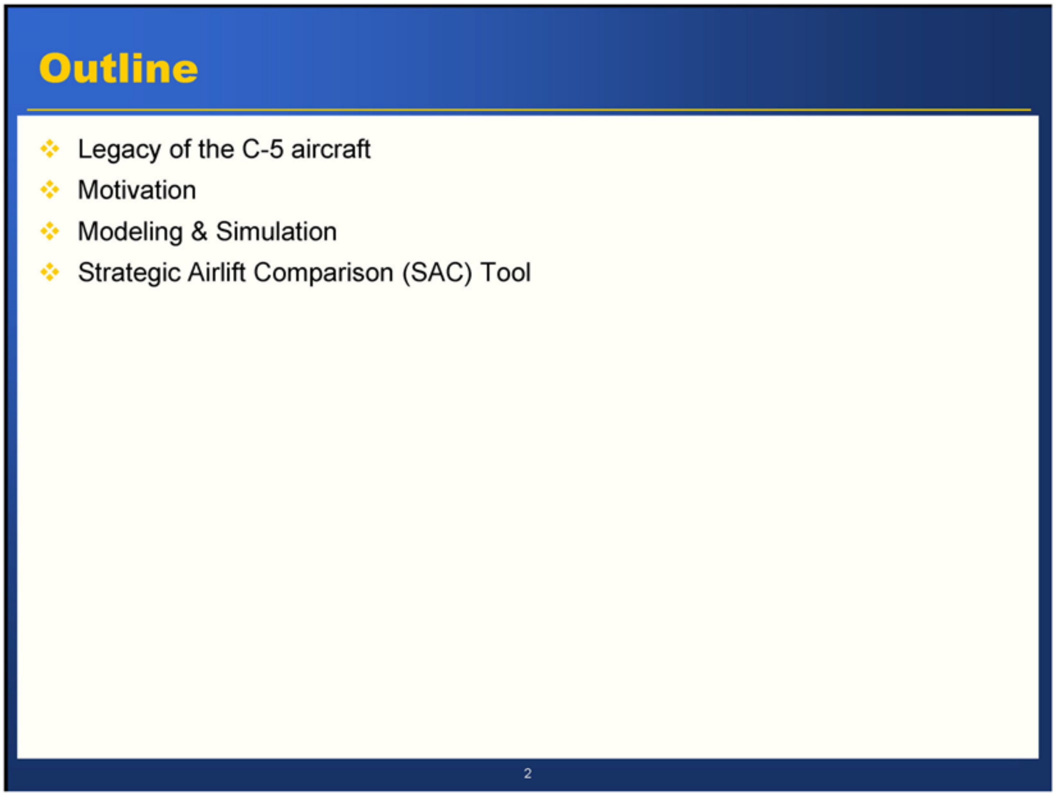
ACKNOWLEDGMENTS

The authors would like to express appreciation for the contributions of Heather Miller, Chris Wyatt, and Walker Armistead for user testing, tool development and Lockheed Martin C-5 performance data, respectively.

Comparative Assessment and Decision Support System for Strategic Military Airlift Capability

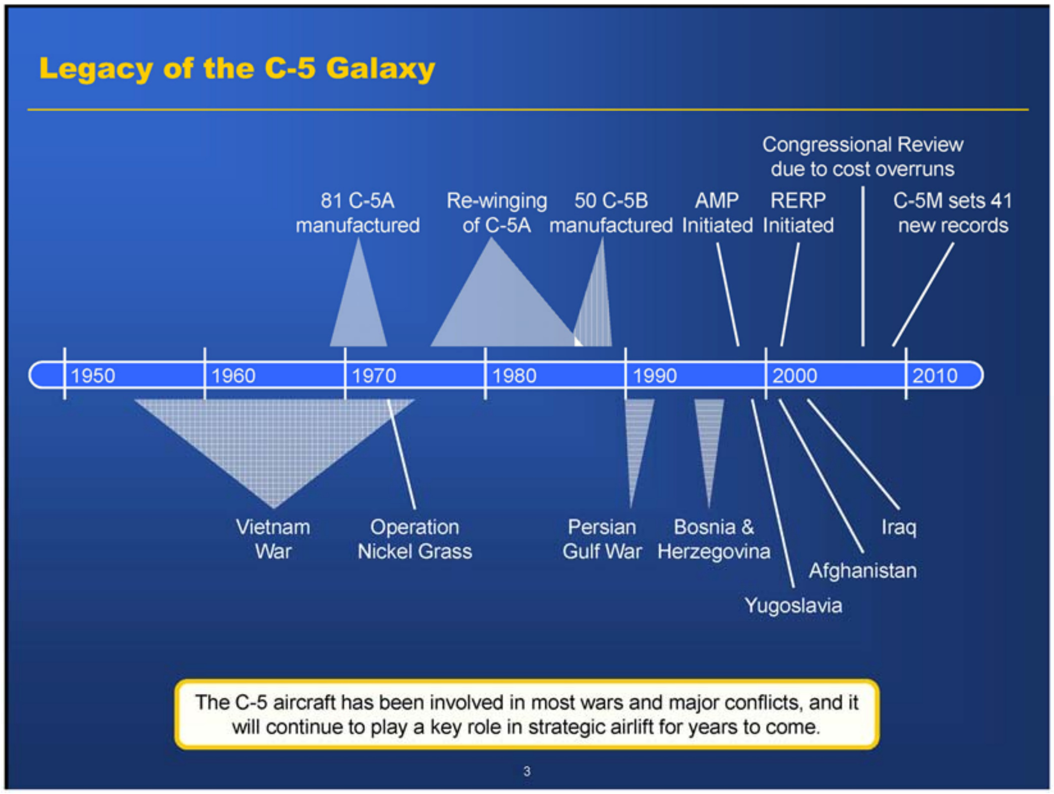
John Salmon
Curtis Iwata
Dimitri Mavris
Neil Weston
Phil Fahringer

Georgia Institute of Technology
ASDL



Outline

- Legacy of the C-5 aircraft
- Motivation
- Modeling & Simulation
- Strategic Airlift Comparison (SAC) Tool



Legacy of the C-5 Galaxy

Congressional Review due to cost overruns

81 C-5A manufactured

Re-winging of C-5A

50 C-5B manufactured

AMP Initiated

RERP Initiated

C-5M sets 41 new records

1950

1960

1970

1980

1990

2000

2010

Vietnam War

Operation Nickel Grass

Persian Gulf War

Bosnia & Herzegovina

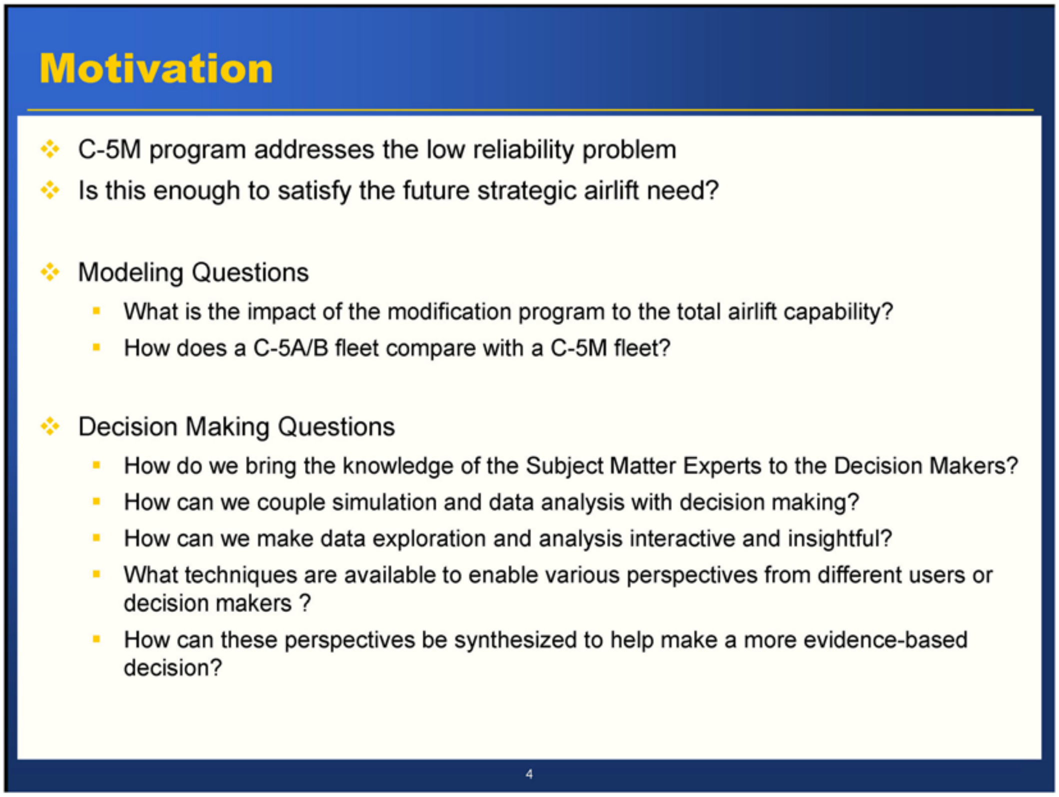
Yugoslavia

Afghanistan

Iraq

The C-5 aircraft has been involved in most wars and major conflicts, and it will continue to play a key role in strategic airlift for years to come.

3



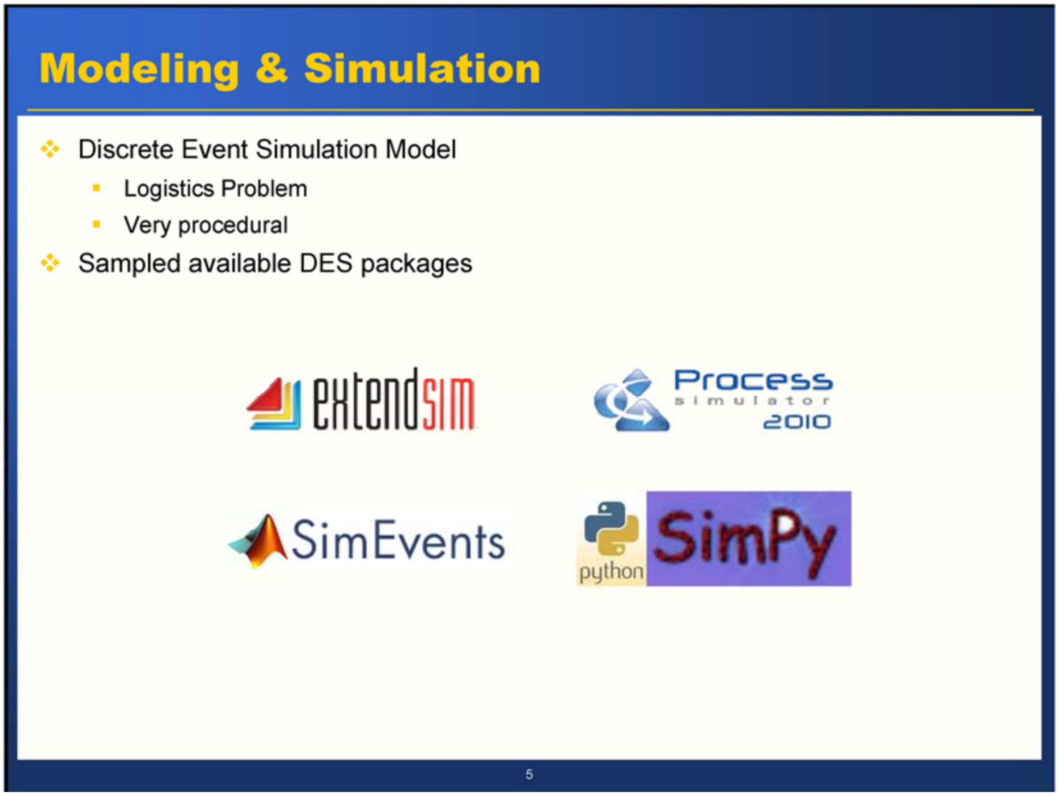
Motivation

- C-5M program addresses the low reliability problem
- Is this enough to satisfy the future strategic airlift need?

- Modeling Questions
  - What is the impact of the modification program to the total airlift capability?
  - How does a C-5A/B fleet compare with a C-5M fleet?

- Decision Making Questions
  - How do we bring the knowledge of the Subject Matter Experts to the Decision Makers?
  - How can we couple simulation and data analysis with decision making?
  - How can we make data exploration and analysis interactive and insightful?
  - What techniques are available to enable various perspectives from different users or decision makers ?
  - How can these perspectives be synthesized to help make a more evidence-based decision?

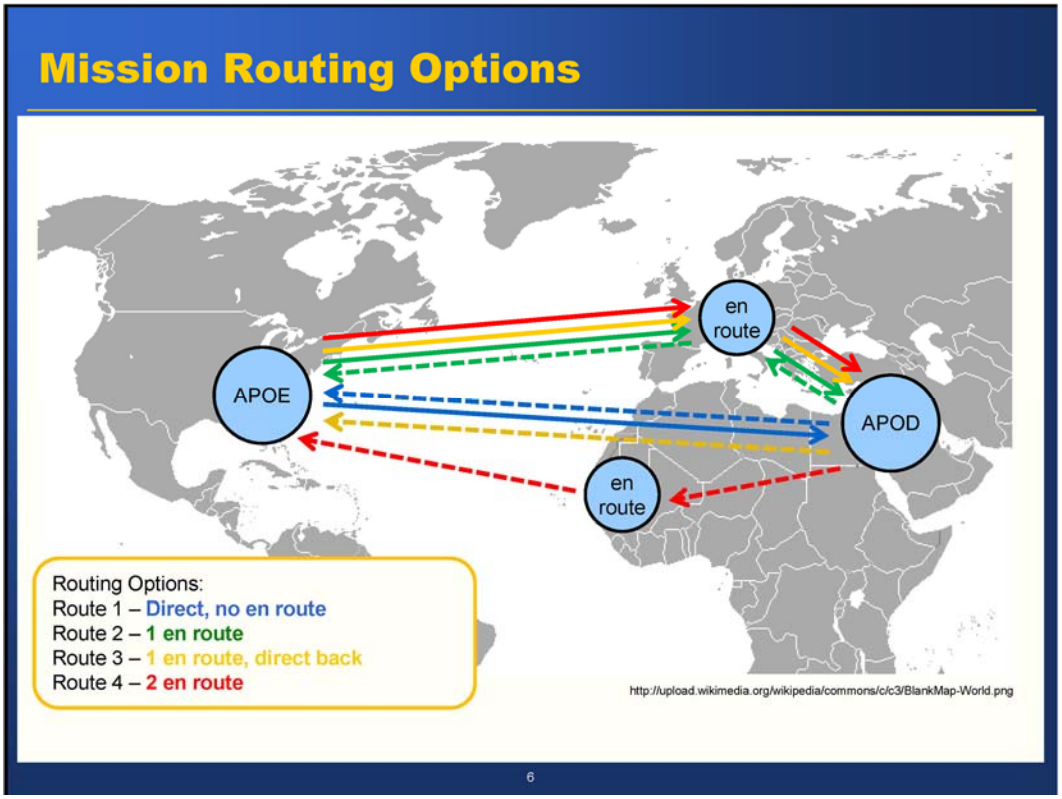
4



Modeling & Simulation

- Discrete Event Simulation Model
  - Logistics Problem
  - Very procedural
- Sampled available DES packages

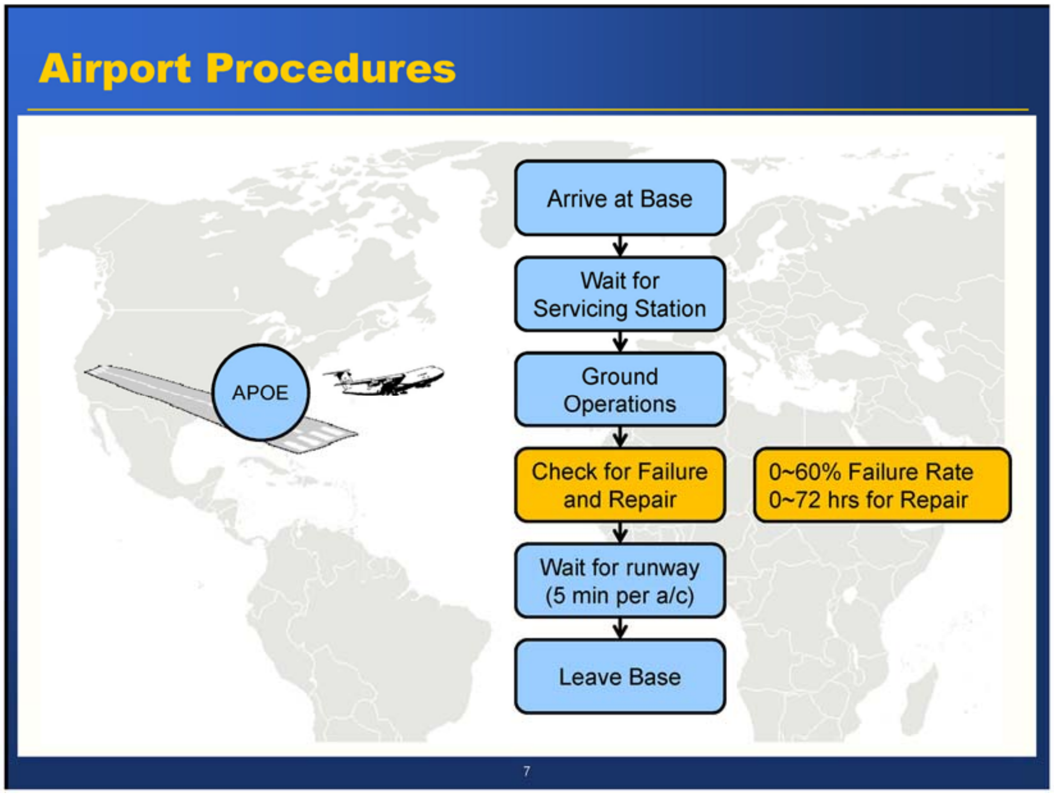
extendsim
Process simulator 2010
SimEvents
python
SimPy



Mission Routing Options

Routing Options:
Route 1 – Direct, no en route
Route 2 – 1 en route
Route 3 – 1 en route, direct back
Route 4 – 2 en route

http://upload.wikimedia.org/wikipedia/commons/c/c3/BlankMap-World.png



Airport Procedures

APOE

Arrive at Base

Wait for Servicing Station

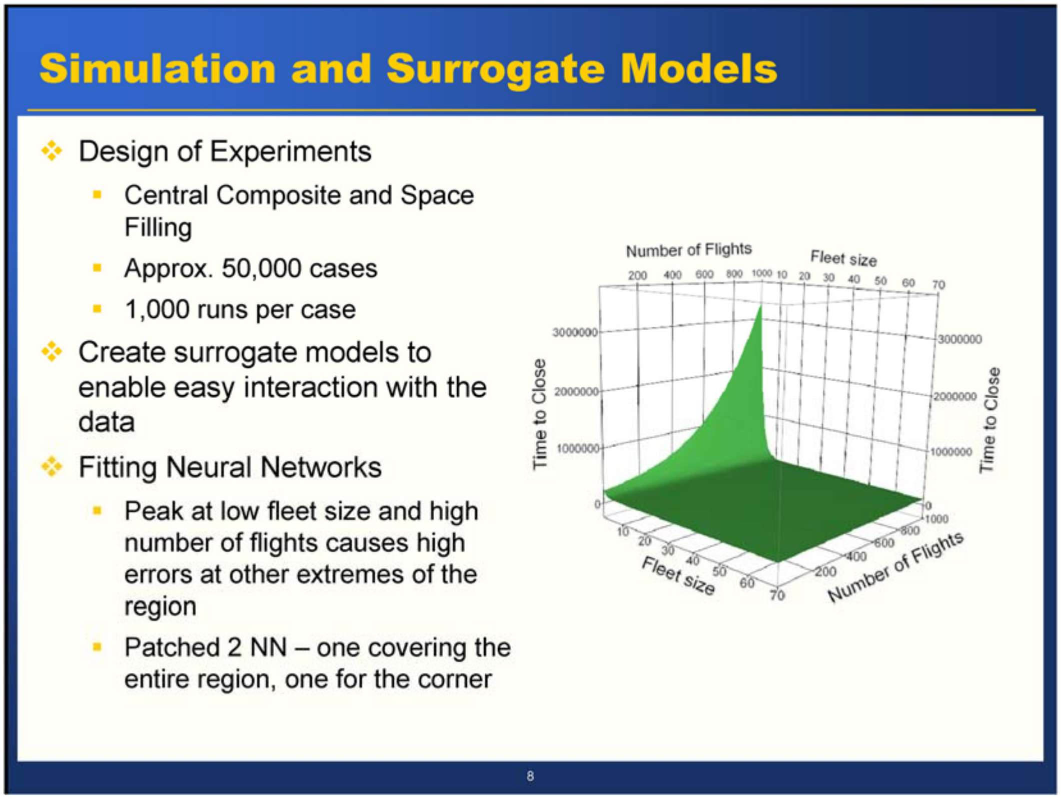
Ground Operations

Check for Failure and Repair

0~60% Failure Rate
0~72 hrs for Repair

Wait for runway (5 min per a/c)

Leave Base



Simulation and Surrogate Models

- Design of Experiments
  - Central Composite and Space Filling
  - Approx. 50,000 cases
  - 1,000 runs per case
- Create surrogate models to enable easy interaction with the data
- Fitting Neural Networks
  - Peak at low fleet size and high number of flights causes high errors at other extremes of the region
  - Patched 2 NN – one covering the entire region, one for the corner

Visualizations and SAC Tool Demo

- Demo

SAC Tool Layout and Analyses

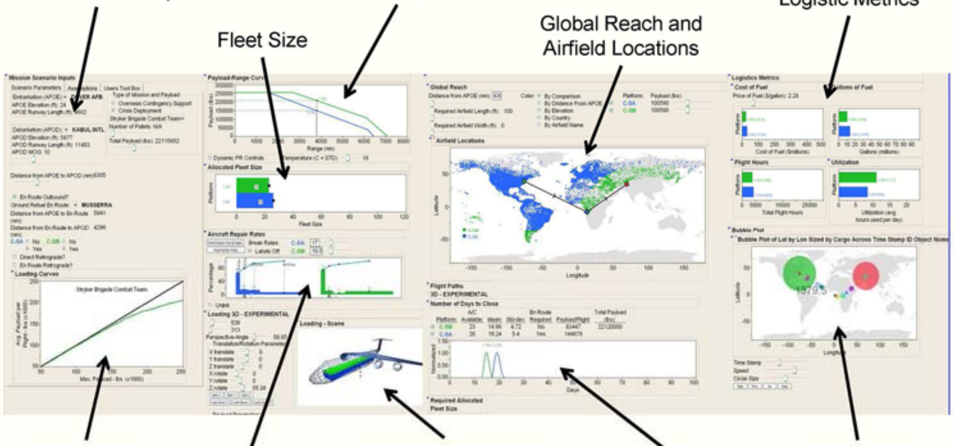
Mission Scenario Inputs

Payload-Range Curve

Fleet Size

Global Reach and Airfield Locations

Logistic Metrics



Loading Curve

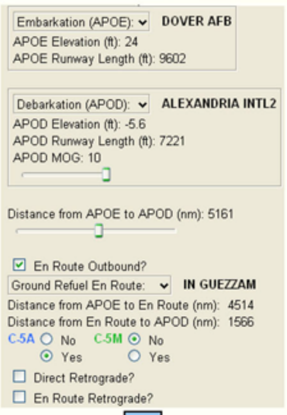
Break and Repair Rates

C-5 Loading Volume

Time to Close

Bubble Plot

Mission Scenario



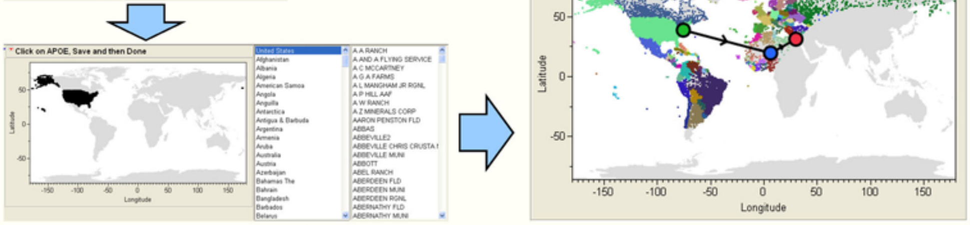
Embarkation (APOE): DOVER AFB
APOE Elevation (ft): 24
APOE Runway Length (ft): 9602

Debarkation (APOD): ALEXANDRIA INTL2
APOD Elevation (ft): -5.6
APOD Runway Length (ft): 7221
APOD MOG: 10

Distance from APOE to APOD (nm): 5161

☑ En Route Outbound?
Ground Refuel En Route: IN GUEZZAM
Distance from APOE to En Route (nm): 4514
Distance from En Route to APOD (nm): 1566
C-5A ○ No ● Yes
C-5M ● No ○ Yes
□ Direct Retrograde?
□ En Route Retrograde?

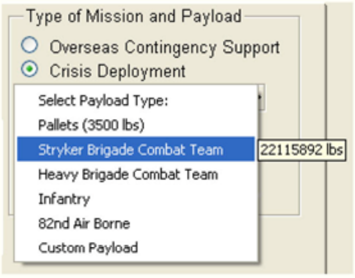
- Dynamic Mission Scenario Selection
  - Aerial Port of Embarkation (APOE), Aerial Port of Debarkation (APOD) and En Route locations selectable from more than 36,000 airfields
  - Unspecified APOD user-defined parameters
  - Adjustable MOG for APOD



Click on APOE, Save and then Done

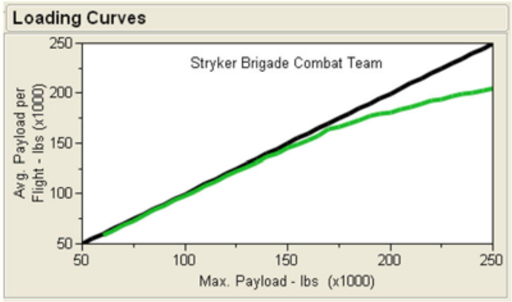
Payload Selection

- Various Selectable Payload types
  - Pallets
  - Stryker Brigade Combat Team
  - Heavy Brigade Combat Team
  - Infantry
  - 82nd Air Borne
  - Custom Payloads
- Cubing-out conditions applied through average payload per flight curves



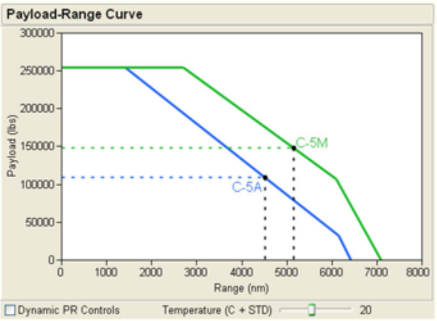
Type of Mission and Payload
○ Overseas Contingency Support
● Crisis Deployment

Select Payload Type: Stryker Brigade Combat Team 22115892 lbs



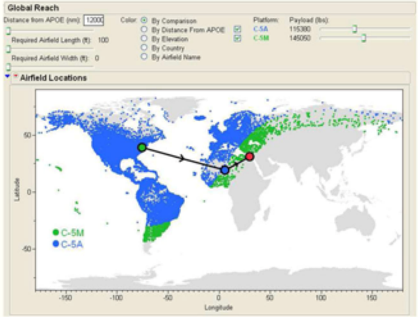
Payload-Range Curve

- Dynamic Payload-Range Curve
  - Comparisons between platforms or investigations in isolation
  - Interactive payload slider bars and operational point
  - Curves for multiple temperatures
  - Reachable airfields based on payload-range curve



□ Dynamic PR Controls

Temperature (C + STD) 20



Distance from APOE (nm): 12000

Required Airfield Length (ft): 100

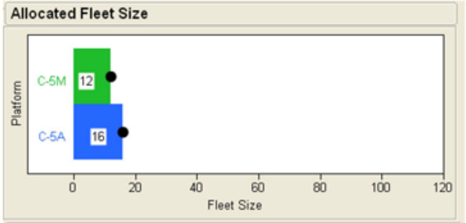
Required Airfield Width (ft): 0

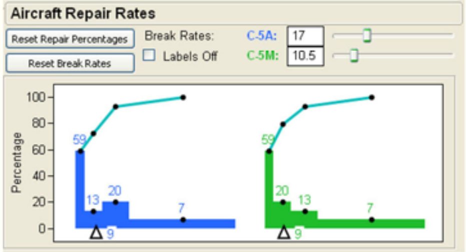
Color: ● By Comparison ○ By Distance From APOE ○ By Elevation ○ By Country ○ By Airfield Name

Platform:	Payload (lbs):
☑ C-5A	115380
☑ C-5M	145050

Allocated Fleet Size and Reliability

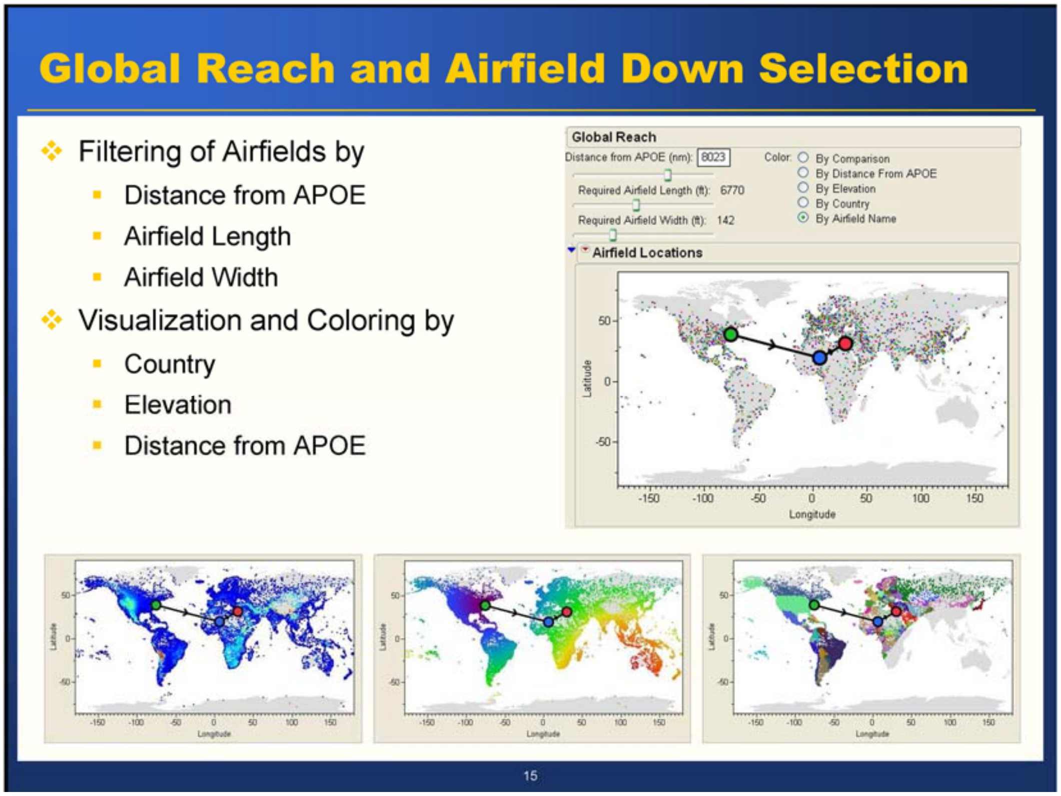
- Fleet Size Comparison
  - Dynamic bar chart
  - Linked to actual number of C-5A and C-5Ms
- Aircraft Reliability Comparisons
  - Repair Rates
  - Break Rates
  - Interactive Probability Density Functions (PDF) and Cumulative Distribution Functions (CDF) for time to repair
  - Adjustable break rates for each platform





Break Rates: C-5A: 17 C-5M: 10.5

□ Labels Off



Global Reach and Airfield Down Selection

- Filtering of Airfields by
  - Distance from APOE
  - Airfield Length
  - Airfield Width
- Visualization and Coloring by
  - Country
  - Elevation
  - Distance from APOE

Global Reach

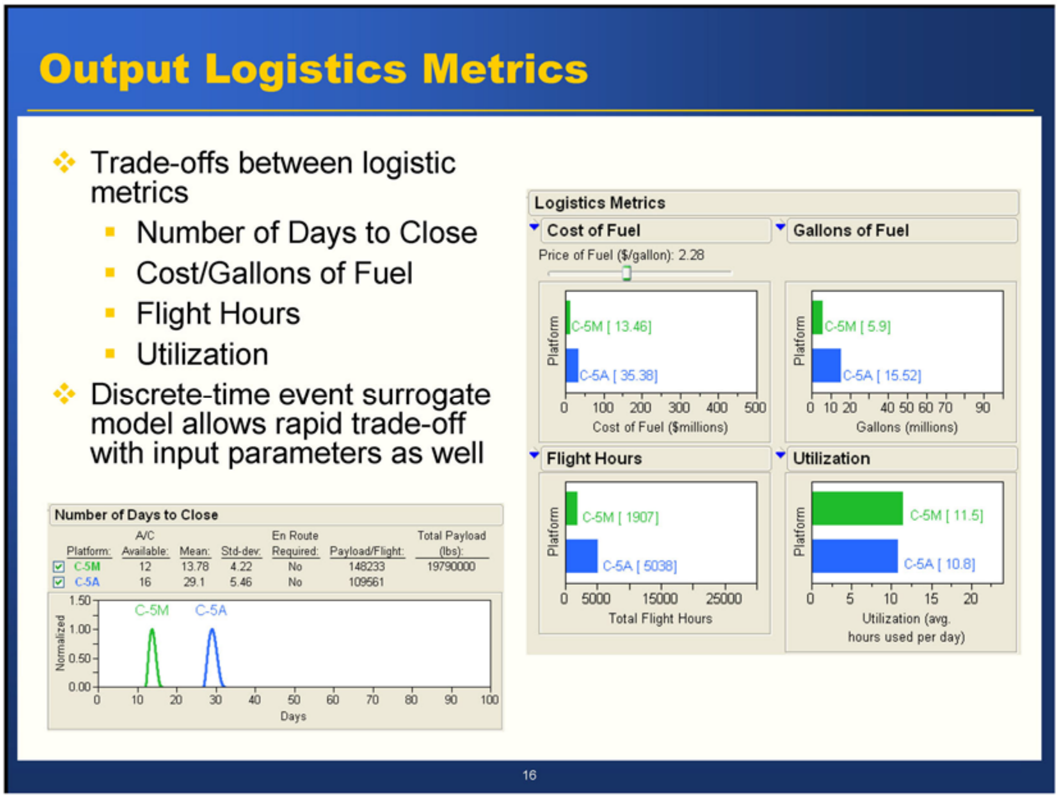
Distance from APOE (nm): 8023

Required Airfield Length (ft): 6770

Required Airfield Width (ft): 142

Color: ○ By Comparison ○ By Distance From APOE ○ By Elevation ○ By Country ● By Airfield Name

Airfield Locations



Output Logistics Metrics

- Trade-offs between logistic metrics
  - Number of Days to Close
  - Cost/Gallons of Fuel
  - Flight Hours
  - Utilization
- Discrete-time event surrogate model allows rapid trade-off with input parameters as well

Number of Days to Close

	Platform:	A/C Available:	Mean:	Std-dev:	En Route Required:	Payload/Flight:	Total Payload (lbs):
☑	C-5M	12	13.78	4.22	No	148233	19790000
☑	C-5A	16	29.1	5.46	No	109561	

Logistics Metrics

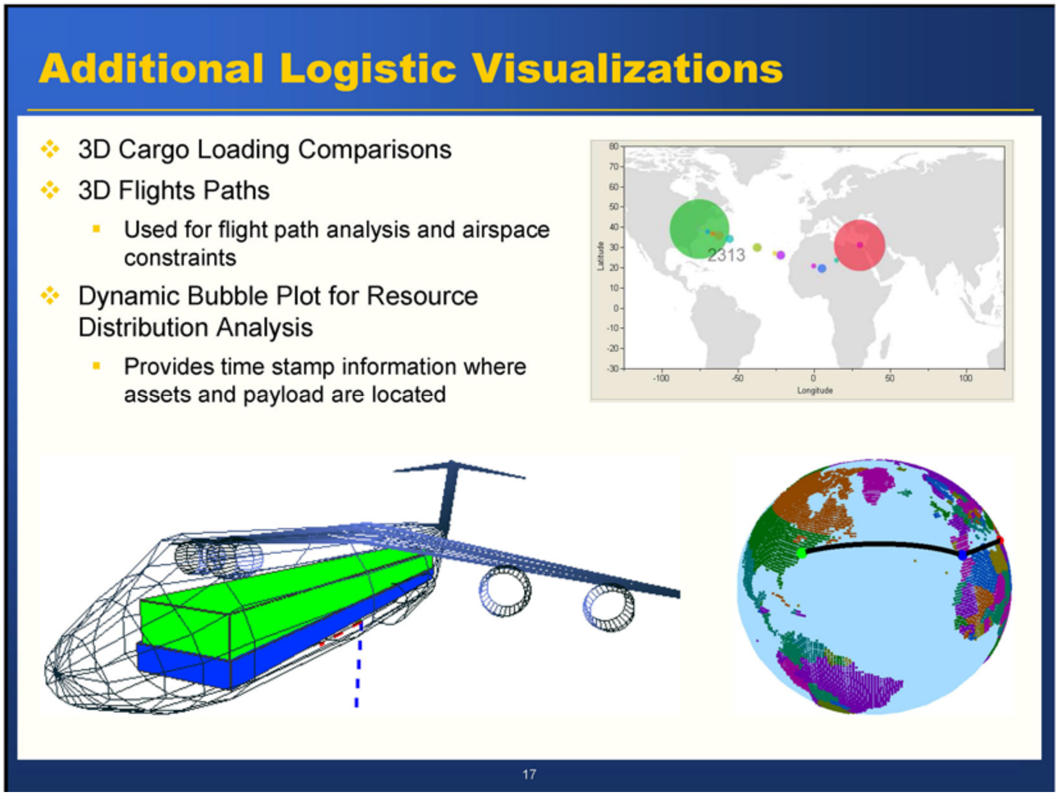
Cost of Fuel

Price of Fuel ($/gallon): 2.28

Gallons of Fuel

Flight Hours

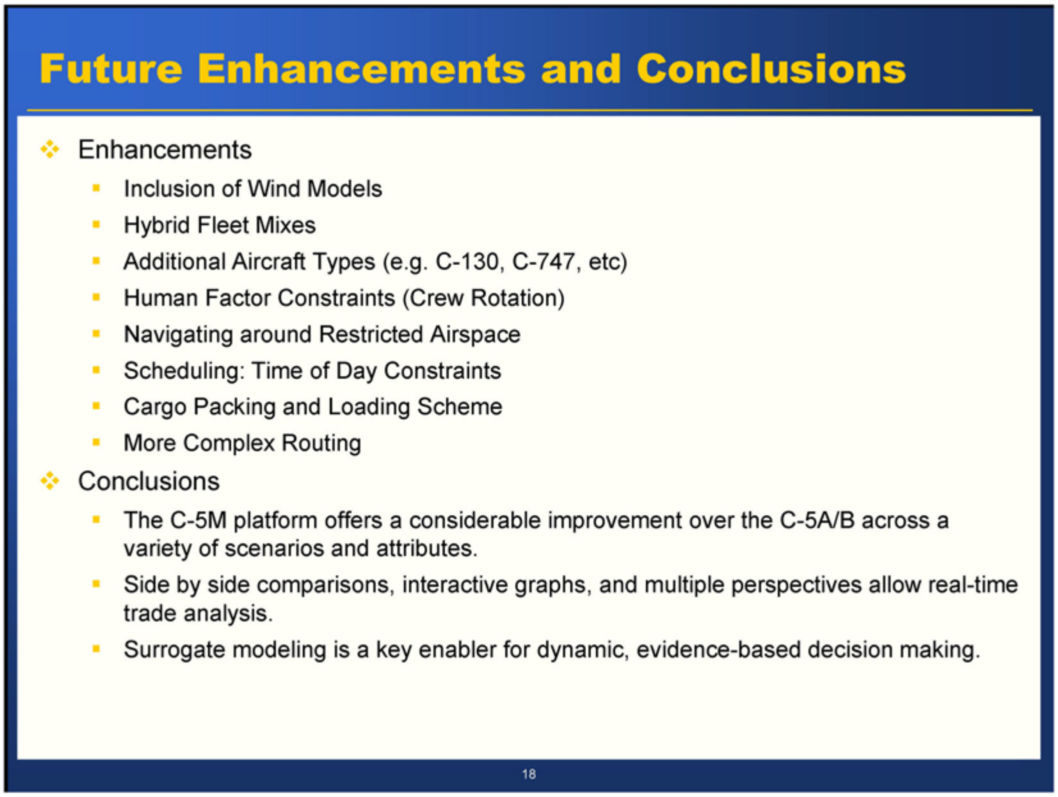
Utilization



Additional Logistic Visualizations

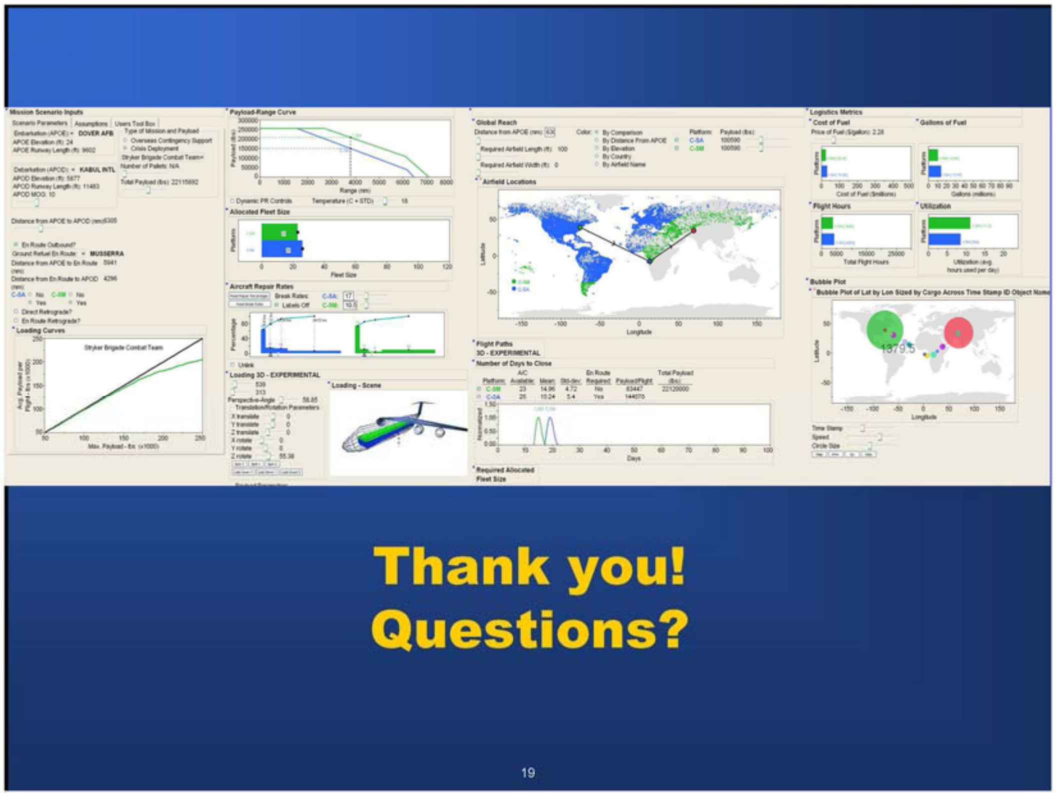
- 3D Cargo Loading Comparisons
- 3D Flights Paths
  - Used for flight path analysis and airspace constraints
- Dynamic Bubble Plot for Resource Distribution Analysis
  - Provides time stamp information where assets and payload are located

80
70
60
50
40
30
20
10
0
-10
-20
-30
Latitude
2313
-100
-50
0
50
100
Longitude



Future Enhancements and Conclusions

- Enhancements
  - Inclusion of Wind Models
  - Hybrid Fleet Mixes
  - Additional Aircraft Types (e.g. C-130, C-747, etc)
  - Human Factor Constraints (Crew Rotation)
  - Navigating around Restricted Airspace
  - Scheduling: Time of Day Constraints
  - Cargo Packing and Loading Scheme
  - More Complex Routing
- Conclusions
  - The C-5M platform offers a considerable improvement over the C-5A/B across a variety of scenarios and attributes.
  - Side by side comparisons, interactive graphs, and multiple perspectives allow real-time trade analysis.
  - Surrogate modeling is a key enabler for dynamic, evidence-based decision making.



Mission Scenario Inputs

Scenario Parameters | Assumptions | Users Tool Box

Embarkation (APOE): DOVER AFB
APOE Elevation (ft): 24
APOE Runway Length (ft): 9602

Debarkation (APOD): KABUL INTL
APOD Elevation (ft): 5877
APOD Runway Length (ft): 11483
APOD MOG: 10

Type of Mission and Payload
○ Overseas Contingency Support
● Crisis Deployment
Stryker Brigade Combat Team
Number of Pallets: N/A
Total Payload (lbs): 22115692

Distance from APOE to APOD (nm) 6305

☑ En Route Outbound?
Ground Refuel En Route: MUSSERRA
Distance from APOE to En Route (nm): 5941
Distance from En Route to APOD (nm): 4296
C-5A ○ No ● Yes
C-5M ○ No ● Yes
□ Direct Retrograde?
□ En Route Retrograde?

Loading Curves

Stryker Brigade Combat Team
Avg. Payload per Flight - lbs (x1000)
250
200
150
100
50
100
150
200
250
Max. Payload - lbs (x1000)

Payload-Range Curve

□ Dynamic PR Controls
Temperature (C + STD): 18

Allocated Fleet Size

Aircraft Repair Rates

Break Rates: C-5A: 17; C-5M: 10.5
☑ Labels Off

□ Unlink

Loading 3D - EXPERIMENTAL

539
313
Perspective-Angle 58.85
Translation/Rotation Parameters
X translate 0
Y translate 0
Z translate 0
X rotate 0
Y rotate 0
Z rotate 55.38

Loading - Scene

Global Reach

Distance from APOE (nm): 63
Required Airfield Length (ft): 100
Required Airfield Width (ft): 0

Color: ● By Comparison
○ By Distance From APOE
○ By Elevation
○ By Country
○ By Airfield Name

Platform:	Payload (lbs):
C-5A	100590
C-5M	100590

Airfield Locations

Flight Paths
3D - EXPERIMENTAL

Number of Days to Close

Platform:	A/C Available:	Mean:	Std-dev:	En Route Required:	Payload/Flight:	Total Payload (lbs):
C-5M	23	14.96	4.72	No	83447	22120000
C-5A	26	19.24	5.4	Yes	144078	

Required Allocated Fleet Size

Logistics Metrics

Cost of Fuel

Price of Fuel ($/gallon): 2.28

Gallons of Fuel

Flight Hours

Utilization

Bubble Plot

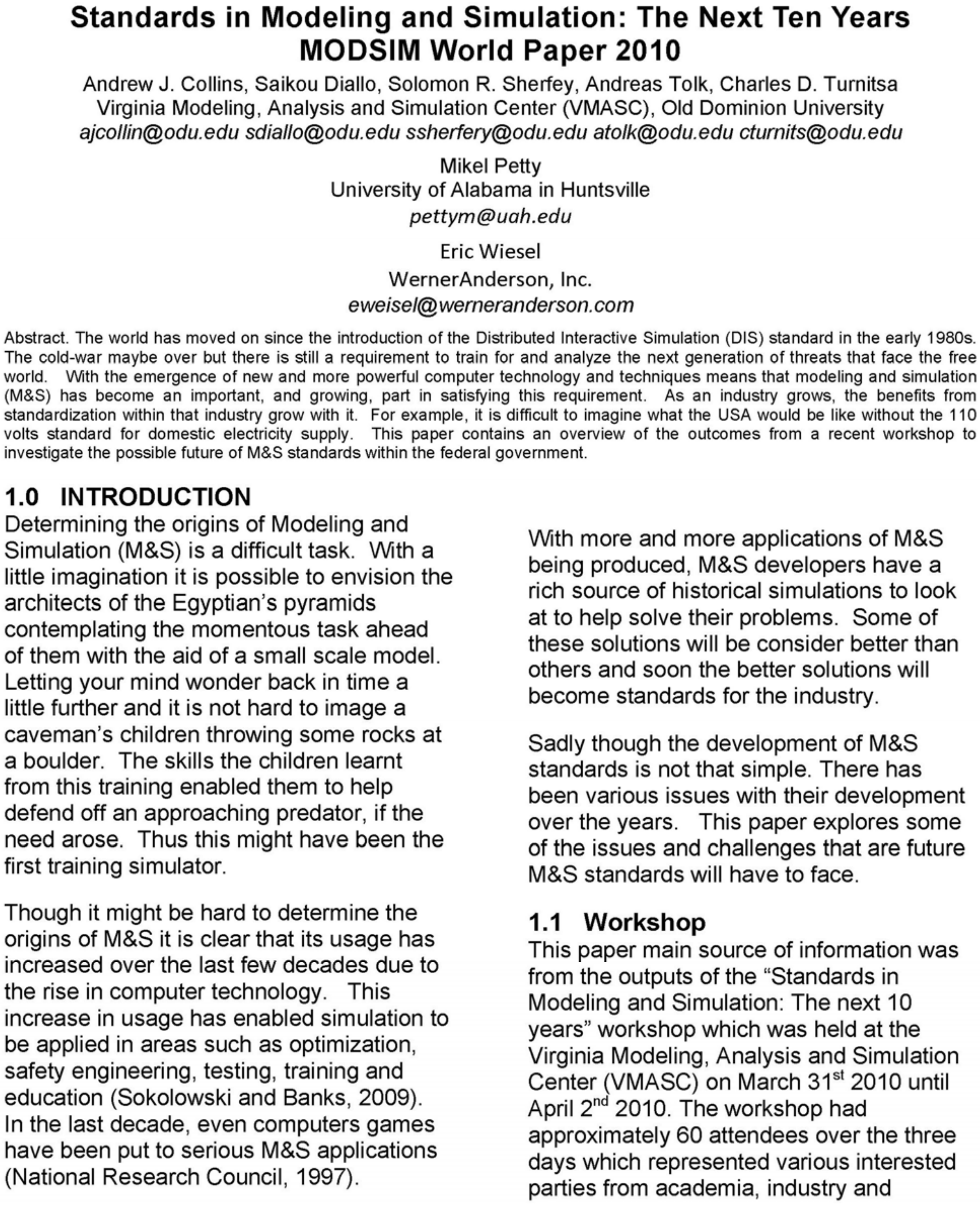
Bubble Plot of Lat by Lon Sized by Cargo Across Time Stamp ID Object Name

Time Stamp
Speed
Circle Size

Thank you!
Questions?

19

1.7 Standards in Modeling and Simulation: The Next Ten Years MODSIM World Paper 2010



Standards in Modeling and Simulation: The Next Ten Years MODSIM World Paper 2010

Andrew J. Collins, Saikou Diallo, Solomon R. Sherfey, Andreas Tolk, Charles D. Turnitsa
Virginia Modeling, Analysis and Simulation Center (VMASC), Old Dominion University
ajcollin@odu.edu sdiallo@odu.edu ssherfery@odu.edu atolk@odu.edu cturnits@odu.edu

Mikel Petty
University of Alabama in Huntsville
pettym@uah.edu

Eric Wiesel
WernerAnderson, Inc.
eweisel@werneranderson.com

Abstract. The world has moved on since the introduction of the Distributed Interactive Simulation (DIS) standard in the early 1980s. The cold-war maybe over but there is still a requirement to train for and analyze the next generation of threats that face the free world. With the emergence of new and more powerful computer technology and techniques means that modeling and simulation (M&S) has become an important, and growing, part in satisfying this requirement. As an industry grows, the benefits from standardization within that industry grow with it. For example, it is difficult to imagine what the USA would be like without the 110 volts standard for domestic electricity supply. This paper contains an overview of the outcomes from a recent workshop to investigate the possible future of M&S standards within the federal government.

1.0 INTRODUCTION

Determining the origins of Modeling and Simulation (M&S) is a difficult task. With a little imagination it is possible to envision the architects of the Egyptian's pyramids contemplating the momentous task ahead of them with the aid of a small scale model. Letting your mind wonder back in time a little further and it is not hard to image a caveman's children throwing some rocks at a boulder. The skills the children learnt from this training enabled them to help defend off an approaching predator, if the need arose. Thus this might have been the first training simulator.

Though it might be hard to determine the origins of M&S it is clear that its usage has increased over the last few decades due to the rise in computer technology. This increase in usage has enabled simulation to be applied in areas such as optimization, safety engineering, testing, training and education (Sokolowski and Banks, 2009). In the last decade, even computers games have been put to serious M&S applications (National Research Council, 1997).

With more and more applications of M&S being produced, M&S developers have a rich source of historical simulations to look at to help solve their problems. Some of these solutions will be consider better than others and soon the better solutions will become standards for the industry.

Sadly though the development of M&S standards is not that simple. There has been various issues with their development over the years. This paper explores some of the issues and challenges that are future M&S standards will have to face.

1.1 Workshop

This paper main source of information was from the outputs of the "Standards in Modeling and Simulation: The next 10 years" workshop which was held at the Virginia Modeling, Analysis and Simulation Center (VMASC) on March 31st 2010 until April 2nd 2010. The workshop had approximately 60 attendees over the three days which represented various interested parties from academia, industry and government. Though the workshop was intended to be focused on military modeling and simulation (M&S) standards, there were individuals from groups outside this arena at the workshop including NASA and the Society for Simulation in Healthcare.

The purpose of the workshop was to give everyone involved an opportunity to think and discuss all aspects of M&S standards over the next 10 years. By conducting the meeting in a non-attributable environment, it allowed participants to engage in more 'out of the box' thinking without being concerned that their ideas would be attributable to themselves or their organization.

1.2 Overview

In the next section of this paper, the terminology and concepts of standards are introduced. Standards specific to M&S are then introduced in section 3.0. The issues and challenges for M&S standards are then discussed in section 4.0. Finally, conclusion to the paper is given in section 5.0.

2.0 STANDARDS

A formal definition of a standard is given by the Federal Office of Management and Budget circular (1998):

> " (1) Common and repeated use of rules, conditions, guidelines or characteristics for products or related processes and production methods, and related management systems practices.
>
> (2) The definition of terms; classification of components; delineation of procedures; specification of dimensions, materials, performance, designs, or operations; measurement of quality and quantity in describing materials, processes, products, systems, services, or practices; test methods and sampling procedures; or descriptions of fit and measurements of size or strength.
>
> The term "standard" does not include the following:
>
> (1) Professional standards of personal conduct.
>
> (2) Institutional codes of ethics."

This definition is no way definitive as there is dissatisfaction within the circle of those who it affects (Finkleman, 2007). As there was some ambiguity to the meaning of standards, the following list gives a more general indication of the different general types of standards.

• **De facto** – standards that have achieved dominant position by public acceptance or market forces i.e. VHS vs. Betamax. A more formal definition of this type of standard was given above in part (1) of the Federal Office of Management and Budget circular (1998).

• **Voluntary** – standards are formally proposed and accepted by a community of interest e.g. key furniture dimensions. In some cases, compliance to a voluntary standard might be necessary to successfully participate in a particular market, for example optional standards requirements on M&S grants.

• **De Dejure** – standards that are mandated by law i.e. residential building codes.

For the purpose of this paper this list is deem adequate definition of standards though the list does not capture every aspect of a standard. For example there are implicit and explicit standards. An example of an implicit standard is a computer game's action configuration on the buttons of a game consul controller, where the user would expect the 'fire' and 'jump' actions to be allocated to certain buttons.

2.1 Why Standards?

The key reason for having M&S standards are: cost savings, technical superiority and convergence. Cost savings come through standards enabling simulation reuse.

Technical superiority is gained by allowing existing technology to work more effectively i.e. simulations federations. Finally, standards allow for convergence of the M&S usage over the different application domains and thus promoting synergy between them.

2.1.1 Requirement for future M&S standards

With the rise of more powerful visualization tools for use within simulation, allowing M&S vendors to give a dazzling display of graphics to potential customers, there have been concerns about a charlatan aspect within the M&S industry. This concern was coined "Garbage in, Hollywood out" in Roman (2005). Thus Validation and Verification M&S standards have been proposed a means to counteract this trend.

There is a requirement for other possible M&S standards, especially relating to data. It should be noted that some data standards already exist (i.e. SEDRIS).

2.2 Standards Bodies

For standards, other than De Facto standards, to come into existence there needs to be an organization that manages their development. For M&S standards, the main body that does this work is the Simulation Interoperability Standards Organization (SISO). SISO was formed in 1989 and became a recognized standards development organization by the IEEE in 2003.

It is not surprising that M&S has its own standards organization, given the vast number of these organizations in the United States (U.S.). Figure 1 gives an overview of all these organizations. To give the reader an idea of scale of this chart, the box that has been circle in the lower-left corner is for Department of Defense standards.

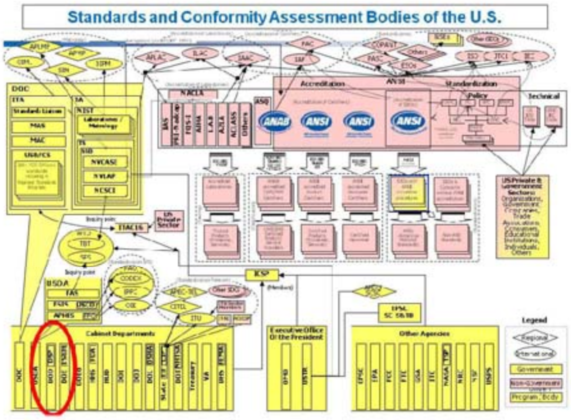


Figure 1: Organizational chart of standards organization with the U.S. (Bipes, 2007)

3.0 M&S STANDARDS

Even when the focus is narrowed from standards to M&S standards, there remain many standards to consider. The list includes both those standards developed specifically for military M&S (e.g., the Test and Training Enabling Architecture, or TENA, for test range-oriented distributed simulation) and those developed for broader applications that have been applied to military M&S (e.g., the Unified Modeling Language, or UML, for conceptual modeling). There are nearly as many governance mechanisms for these standards as there are standards, ranging from the highly formal (e.g., IEEE standards with explicitly defined community voting procedures for standards revisions) to the highly informal (e.g., some military standards with appointed panels of users, technical experts, and sponsors deciding on proposed revisions).

Beyond governance formality, M&S standards also vary by degree of technical specificity, defined as the extent to which a particular technical utilization or solution is mandated by the standard. In this attribute the standards range from very low specificity, such as broad guidelines to users (e.g., diagram formats in UML), to very high technical specificity, such as common software components required for all users (e.g., interoperability middleware in TENA). It has been conjectured that these attributes of standards, governance formality and technical specificity, perhaps combined with other attributes can be shown to be correlated with or even predictive of the expected utility and ubiquity of standards for military M&S.

The DoD's Modeling and Simulation Coordination Office (MSCO) is currently engaged in a comprehensive survey of military M&S standards. An outcome of the survey will be the cataloging and characterization of military M&S standards on many attributes of interest. This effort will enable standards developers and users to understand what standards are available to support ongoing and planned military M&S development projects and to identify gaps where new standards might be beneficial. The mechanisms by which standards in general, and military M&S standards in particular are created and maintained vary widely. An understanding of the processes available can guide the selection of the appropriate venue and process for introducing a new proposed standard.

One of the most important and successful M&S standards of recent years is the High Level Architecture (HLA), which was derived from the Distributed Interactive Simulation (DIS) standard during the early nineties (Hollenbach, 2009).

4.0 DISCUSSION

During the workshop, several topic areas relating to standards were discussed. In this section, a review of some of the key topics is given.

4.1 Measuring M&S standards

The idea of considering standards success as a research question, and associating that success with specific standards attributes, was well received. Return on Investment (ROI) has been the type of measure of a standard's worth due to its implications on cost, as opposed to quality and reliability improvements. However, it is hard to define ROI for an M&S standard as it is difficult to know the impact if the standard was not implemented, therefore ROI should not be the only focus of a standards worth.

Other possible measures include the value that a standard brings to a simulation, the technology advancement it enables, the project risk it mitigates and the impact on the M&S community.

4.2 Confusion

Within the M&S world, there is a lot of confusion relating the terminology used. For instance, the terms 'modeling' and 'simulation' are often used interchangeably even though they are distinct concepts. Some definitions for these terms is given below:

- **Model** - A physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, or process. [DOD, 1998]

- **Simulation** - A method for implementing a model over time. Also, a technique for testing, analysis, or training in which real world systems are used, or where a model reproduces real world and conceptual systems. [DOD, 1998]

Other terms that are confused include Fidelity (the accuracy of model's representation or simulation's results) and resolution (the degree of detail with which the real-world is simulated). An example of a high resolution but low fidelity simulation would be Microsoft flight simulator.

Another example is composability and interoperability. Composability is different from interoperability as it implies models working together to produce a valid whole meta-model. Interoperability simply implies that two simulations are able to communicate with each other.

To understand this difference, consider a 'fish tank' simulation and a combat simulation. Both simulations might have an object within them called 'tank', and they might be able to interoperate by passing the ‘tank’ object to each other. However, the concept of the ‘tank’ object is completely different in each model and thus the models are not composable.

The challenge of composability brings to M&S are some of the most difficult ones that M&S research have faced in recent years.

4.3 Education

The complexity of M&S is likely to increase in the coming years and this will increase the need for more education of M&S and standards as well. Given the confusion associated with M&S terminology, this education should apply to customers of M&S as well as the next generation of professionals.

It is, however, important to remember that users are not interested in M&S standards; they are interested in the functionality that happens because of standards.

Standards themselves are not a new concept and they already impact on every aspect of our lives. There was no consensus within the workshop whether it was appropriate to compare M&S standards to other existing standards though several useful standards analogies were given. It was clear that standards do have a lifecycle and it was suggested that we should focus our efforts on the standards that are likely to have the longest lifecycles. This focus might help mitigate some of the ‘lag effect’ that standards tend to have compared to cutting edge technology.

5.0 CONCLUSION

The future of M&S will see a changing shift from construction simulation to live and virtual simulation. With this shift come lots of standards requirements from new problems and areas that are appearing i.e. VV&A standards and data standards. The future also holds many challenges for M&S including composability and determining the ROI from M&S standards.

Though M&S is an emerging new discipline, stagnation of its development can and has occurred. Therefore, it is difficult to say what M&S will be like in 10 years, it might be similar to the situation of today or a completely unimaginable area.

The workshop did not cover all aspects of standards and there are many more discussion areas that need to be addressed. The next workshop in the series was held on, which will be on M&S standards governance, was held on August 4th to 6th 2010.

Return your imagination back to the caveman’s children and their throwing of stones at the boulder, which was discussed at the start of this paper. After a while of the playing this game, the children would, no doubt, impose a limit on how close you were allowed to get to the boulder before throwing stone because otherwise the game would become too easy. This would have been the first M&S standard to be developed as it is part of human nature to organize things. Thus the question of future M&S standards should not be about ‘if’ they are going to occur but instead it should be about ‘when they will occur’, ‘by whom’ and ‘how’.

6.0 REFERENCES

[1] Sokolowski, J.A., and Banks, C.M. (2009); “Principles of Modeling and Simulation”; John Wiley and Sons.

[2] National Research Council (1997); “Modeling and Simulation: Linking Entertainment and Defense”; National Academic Press.

[3] Office of Management and Budget (1998); “Federal Participation in the development and use of voluntary consensus standards in conformity assessment activities”; Circular No. A-119 Revised; www.whitehouse.gov/omb/rewrite/circulars/a119/a119.html (accessed on May 26th, 2010).

[4] Finkleman, D. (2007); “A call to action”; Aerospace America 45, no. 11.

[5] Roman, P.A. (2005); "Garbage in, Hollywood out!"; In SimTecT 2005, SimTecT, Sydney, Australia.

[6] Steven Bipes (2007); "Standards and Conformity Assessment Bodies of the United States"; American National Standards Institute (ANSI) Public Document Library, version 2006-07-21.

[7] Hollenbach, J.W. (2009); "Inconsistency, Neglect, and Confusion: A Historical Review of DoD Distributed Simulation Architecture Policies"; In Joint 2009 Spring Simulation Interoperability Workshop (SIW), Joint 2009 Spring Simulation Interoperability Workshop (SIW), San Diego-Mission Valley, CA: Simulation Interoperability Standards Organization.

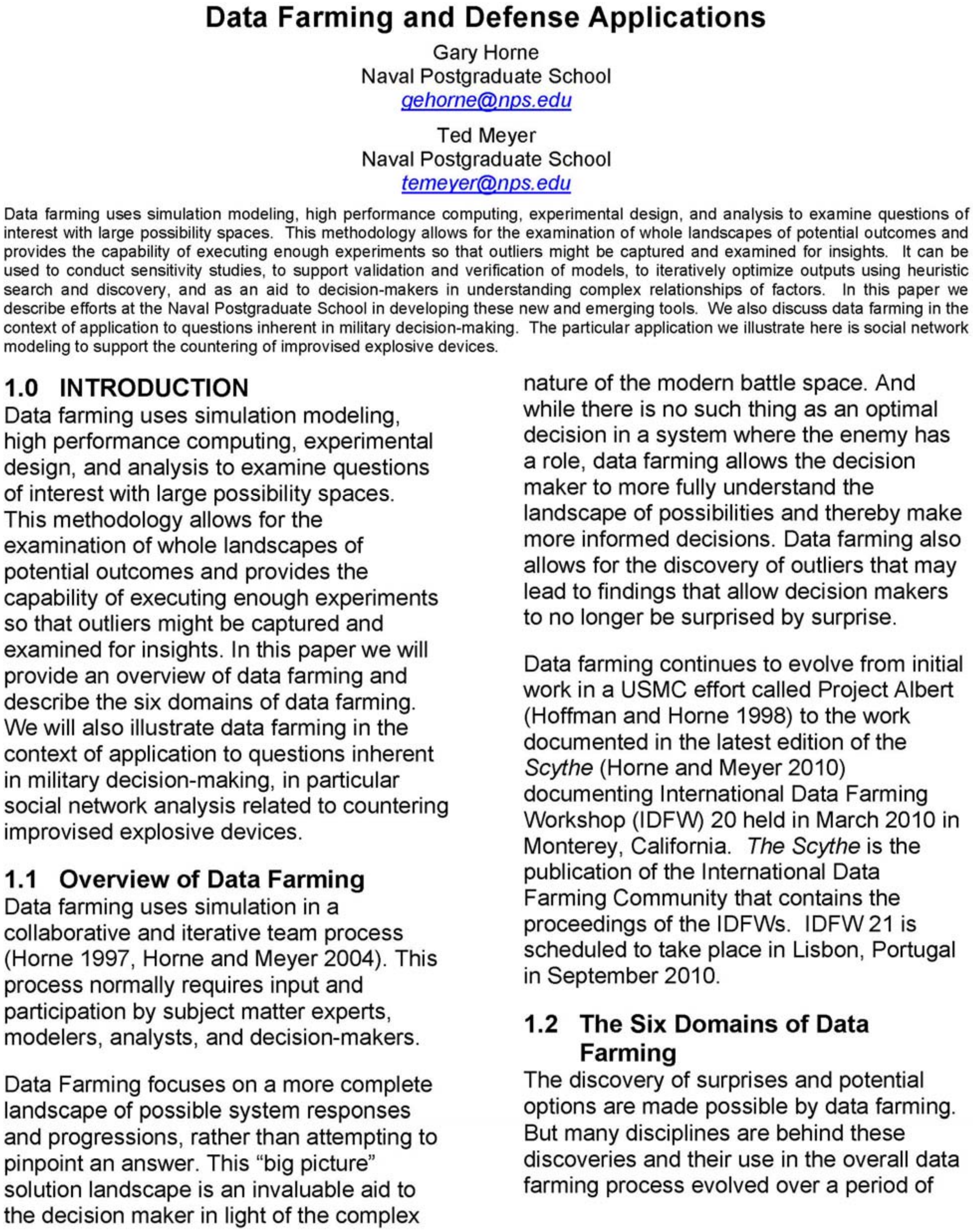
[8] Department of Defence (1998);, "DoD Modeling and Simulation (M&S) Glossary"; DoD Directive 5000.59-M.

7.0 ACKNOWLEDGMENT(S)

Many thanks must be given to all the workshop participants. Without their wisdom and insights, there would be nothing to report in this paper. As participants were told that the workshop was non-attainable, it would be inappropriate to give individual credit and thanks though there are many how deserve it.

The workshop was conducted as part of a federally funded three year study entitled 'Standards in Modeling and Simulation', which is being lead by VMASC.

1.8 Data Farming and Defense Applications



Data Farming and Defense Applications

Gary Horne
Naval Postgraduate School
gehorne@nps.edu

Ted Meyer
Naval Postgraduate School
temeyer@nps.edu

Data farming uses simulation modeling, high performance computing, experimental design, and analysis to examine questions of interest with large possibility spaces. This methodology allows for the examination of whole landscapes of potential outcomes and provides the capability of executing enough experiments so that outliers might be captured and examined for insights. It can be used to conduct sensitivity studies, to support validation and verification of models, to iteratively optimize outputs using heuristic search and discovery, and as an aid to decision-makers in understanding complex relationships of factors. In this paper we describe efforts at the Naval Postgraduate School in developing these new and emerging tools. We also discuss data farming in the context of application to questions inherent in military decision-making. The particular application we illustrate here is social network modeling to support the countering of improvised explosive devices.

1.0 INTRODUCTION

Data farming uses simulation modeling, high performance computing, experimental design, and analysis to examine questions of interest with large possibility spaces. This methodology allows for the examination of whole landscapes of potential outcomes and provides the capability of executing enough experiments so that outliers might be captured and examined for insights. In this paper we will provide an overview of data farming and describe the six domains of data farming. We will also illustrate data farming in the context of application to questions inherent in military decision-making, in particular social network analysis related to countering improvised explosive devices.

1.1 Overview of Data Farming

Data farming uses simulation in a collaborative and iterative team process (Horne 1997, Horne and Meyer 2004). This process normally requires input and participation by subject matter experts, modelers, analysts, and decision-makers.

Data Farming focuses on a more complete landscape of possible system responses and progressions, rather than attempting to pinpoint an answer. This "big picture" solution landscape is an invaluable aid to the decision maker in light of the complex nature of the modern battle space. And while there is no such thing as an optimal decision in a system where the enemy has a role, data farming allows the decision maker to more fully understand the landscape of possibilities and thereby make more informed decisions. Data farming also allows for the discovery of outliers that may lead to findings that allow decision makers to no longer be surprised by surprise.

Data farming continues to evolve from initial work in a USMC effort called Project Albert (Hoffman and Horne 1998) to the work documented in the latest edition of the *Scythe* (Horne and Meyer 2010) documenting International Data Farming Workshop (IDFW) 20 held in March 2010 in Monterey, California. *The Scythe* is the publication of the International Data Farming Community that contains the proceedings of the IDFWs. IDFW 21 is scheduled to take place in Lisbon, Portugal in September 2010.

1.2 The Six Domains of Data Farming

The discovery of surprises and potential options are made possible by data farming. But many disciplines are behind these discoveries and their use in the overall data farming process evolved over a period of time. In this section we give a brief account of this development.

Six realms or domains were incorporated into the data farming methodology from 1997 to 2002. Initial data farming efforts in the 1997-98 time frame relied upon two basic ideas:

1. Developing models, called distillations, which may not have a great deal of verisimilitude but could be focused to specifically address the questions at hand. (Horne 1999)
2. Using high performance computing to execute models many times over varied initial conditions to gain understanding of the possible outliers, trends, and distribution of results

The models need not be agent-based models, but because of the ease with which they can be prototyped, agent-based models were used during this beginning time period. This rapid prototyping facilitated the iterative nature of the approach the use of high performance computing to execute models many times over varied initial conditions to gain understanding of the possible outliers, trends, and distribution of results. Also, the huge volume of output from the simulations made possible by the high performance computing resulted in a need to develop visualization tools and methods commensurate with this tremendous amount of data. Thus, visualization of simulation data and rapid prototyping of scenarios became important to data farming efforts in the 1999-2000 time frame.

The simulations that defense analysts use are often large and complex. An evaluation of complete landscapes is extremely time consuming, sometimes not even possible. Also, even the smaller more abstract agent-based distillations referred to above can have many parameters that are potentially significant and that could take on many values. Thus, even with high performance computing and the small models used in data farming, gridded designs, where every value is simulated, are unwieldy.

Thus, using efficient experimental designs is essential and The Naval Postgraduate School in Monterey, California joined Project Albert researchers in the early 2000s with their expertise in this area. And NPS researchers have collaborated with others worldwide as well (see Kleijman, Sanchez, Lucas, and Cioppa 2005).

Finally, collaboration must take place at many levels if the full power of data farming is to be brought upon any question. Collaborative processes help to integrate the other five domains of data farming through interdisciplinary work in creating models and data farming infrastructure and during the iterative process of prototyping scenarios and examining output from model runs. Collaboration also takes place between people from different organizations and nations sharing information and perspectives at various points in approaching common questions.

With the addition of design of experiments and collaborative processes in 2001-2002 to data farming efforts, much attention then focused on the defense applications discussed in the next section. The six realms, or domains, discussed above that contribute to the data farming process are depicted in Figure 1.

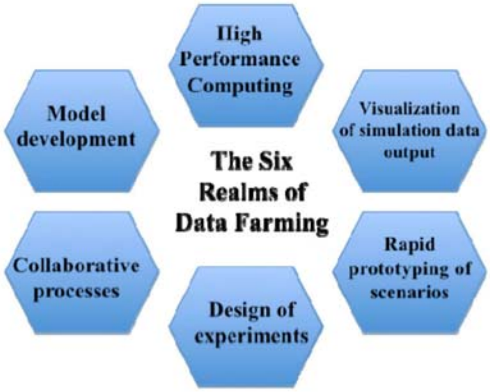

High Performance Computing
Model development
Visualization of simulation data output
The Six Realms of Data Farming
Collaborative processes
Rapid prototyping of scenarios
Design of experiments

Figure 1. The Six Domains of Data Farming

1.3 Defense Applications

Since the incorporation of the above six domains into the process we call data farming, several articles have captured the fundamentals of data farming (e.g. Horne and Meyer 2005). But the key tenet in the data farming process has been the focus on the questions and since 2002 many application efforts have been documented. For example, at the Naval Postgraduate School many theses have been completed which have used data farming. And over the past decade, over 150 international work teams have formed around questions at International Data Farming Workshops.

These 150 work teams fall into areas, or themes, which include: Joint and Combined Operations (e.g. C4ISR Operations, Network Centric Warfare, Networked Fires, and Future Combat Missions), Urban Operations, Combat Support (e.g. UAV Operations, Robotics, Logistics, and Combat ID), Peace Support Operations, the Global War on Terrorism, Homeland Defense, Disaster Relief, and others.

The types of questions in these areas typically do not have precisely defined initial conditions and a complete set of algorithms that describe the system being considered. These questions address open systems that defy prediction. Data farming is used to provide insight that can be used by decision-makers. As an illustrative example, we now describe how data farming is being integrated with other techniques in the context of countering improvised explosive devices.

2.0 ILLUSTRATIVE APPLICATION: SOCIAL NETWORK MODELING TO SUPPORT THE COUNTER-IED FIGHT

This work represents results from an ongoing study to examine the utility of distillation modeling in the Counter-IED (Improvised Explosive Devices) fight. Understanding social networks, their nature in insurgencies and IED networks, and how to impact them, is important to the Counter-IED (C-IED) fight. This study, conducted as a team effort with international and inter-agency participation, is exploring methods of extracting, analyzing, and visualizing dynamic social networks that are inherent in models with agent interaction. This effort is being conducted in order to build tools that may be useful in examining and potentially manipulating insurgencies. The team started with a simple scenario that evolves cliques via interactions based on shared attributes. This simple model is the initial basis for the team's investigations and is being used to examine the types of network statistics that can be used as MOEs and pointers to unique and emergent behaviors of interest.

The team's initial goals were to extend this very basic scenario with simple variations and to test candidate tools and prototype methods for data farming the scenario, extracting network data, analyzing end-of-run network statistics, and visualizing network behaviors.

Social Network Analysis (SNA) techniques were explored in detail to determine which network metrics would be most beneficial for analyzing the types of networks produced by the agent-based scenario. Developing these tools and methods, and delineating applicable metrics will allow the exploration of questions regarding C-IED issues—including insurgent network evolution and adaptation.

Insurgent networks can be categorized into two groups of interest to C-IED efforts: IED Emplacement Networks (consisting of personnel that are directly involved with IED usage) and IED Enabling Networks (consisting of communities that indirectly support and enable the IED Emplacement networks). This study is identifying tools that can be used to explore patterns that might provide valuable insights into emergent behaviors of interest for both of these classes of networks.

2.1 Background

In previous work related to the use of agent-based modeling in the C-IED work, task plans aimed at addressing specific C-IED questions were developed. The current work is aimed at producing capabilities that can address these tasks. Tasks topics included: methods of indirect network attack; identifying important link layers for impacting the insurgent networks in specific environments, identifying important individuals, emergence of insurgent cells, eroding popular support for insurgent networks.

From this set of tasks the study team selected a set of candidate tasks for follow–up study and analysis. It was concluded that both data farming and SNA concepts and techniques needed to be applied to address the candidate tasks and that the current set of tools and methods available in these domains was not up to the task required.

The study team is working on developing the necessary tools and methods. In this effort we have:

- Demonstrated the ability to extract social network data from an existing scenario that included agent interaction, but that did not explicitly define a network. In this scenario the network "emerged" or evolved from the basic agent interactions.

- Data farmed this initial scenario and established the need to simplify the target scenario in order to more closely examine cause and effect relationships to SNA statistics.

- Developed a new base scenario, delineated a simple illustrative design of experiment (DOE), and data farmed the model to provide a sample data set for further exploration.

- Examined the utility of and approach to applying specific SNA statistics, methods, and concepts using the data farming output provided from previous work.

- Delineated the data requirements for the various types of networks that might be extracted from various modeling.

- Established and documented software and processes for applying these capabilities to detecting and analyzing emergent networks.

This work has lead to the study team's conviction that additional work needs to be accomplished in order to address C-IED-oriented problems. Generalized SNA/data farming tools that can be applied to output from various model types should have the capability to:

- Detect the presence of a network or networks.

- Distinguish different networks and different classes of networks.

- Determine if and when networks achieve equilibrium.

- Determine which model inputs have significant impact on the state and behaviors of the network.

Specifically, the intent is to use these capabilities to be able to address a variety of social network questions such as:

- What do insurgent networks look like? Who is in the network? Who is not?

- How do we distinguish networks that should be attacked, networks that should be attritted or that should be co-opted?

- Who are the High Value Individuals (HVI) and what are their identifiable characteristics?

- Will removing specific nodes destabilize a network?

- What are the 2nd and 3rd order effects of network manipulation?

- What are the potential unintended consequences?

2.2 Abstracted Illustrative Scenario and DOE

The Pythagoras agent-based model development environment was used for the initial scenarios. The first phase of activity was based on the Pythagoras distribution "Peace" scenario with some minor modifications of the source code to support the extraction of network interaction data. Data farming of this scenario demonstrated the ability to extract inherent emergent network data. Initial analysis of the results led to the development of a more basic scenario in order to test basic network concepts.

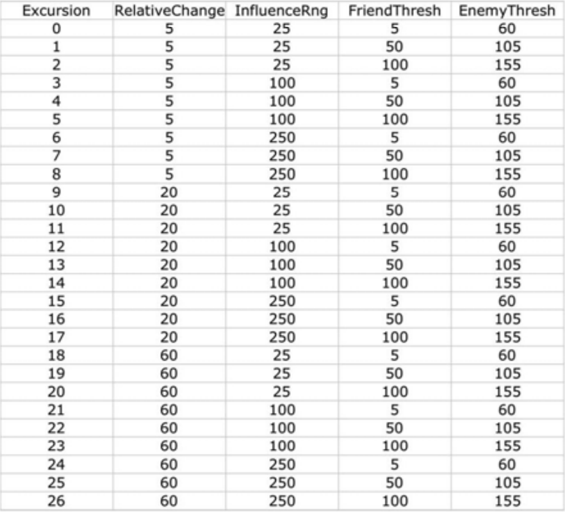
The illustrative "Clique Creator" (CC) scenario was developed using Pythagoras's "relative" color change capability as a tool for experimenting with SNA extraction and analysis. CC has a single agent class with 100 instantiated agents that are uniformly distributed across Pythagoras's red and blue color spaces. The agents' only "weapon" is "Chat" which induces a relative color change on other agents with which the agent interacts. As the scenario is executed, entities move through various color states, becoming "more" red or "more" blue depending on the interactions with other red-"ish" or blue-"ish" entities. States will change depending on whether two entities engage in "chatting" and form a connection. The more any two agents interact, the more "alike" they become.

The focus of the scenario selection was to represent dynamic homophily and use the results to explore the various analysis tools under study. Multiple excursions and replications of the Pythagoras-developed Clique Creator scenario were used to produce the data for analysis with the candidate tools. This baseline provided a means for the team to experiment with various SNA measures and analysis techniques.

Pythagoras can provide multiple views of agent state data. A spatial view showed the physical relationship between entitities and where connections or bonds were formed. The inclination space view sorted the entities by colors. This color space view is used to illustrate the homophilic state of the participating entities in the simulation.

A very basic full-factorial design space was used to data farm the scenario.

Table 1. Experimental Design Matrix



Excursion	RelativeChange	InfluenceRng	FriendThresh	EnemyThresh
0	5	25	5	60
1	5	25	50	105
2	5	25	100	155
3	5	100	5	60
4	5	100	50	105
5	5	100	100	155
6	5	250	5	60
7	5	250	50	105
8	5	250	100	155
9	20	25	5	60
10	20	25	50	105
11	20	25	100	155
12	20	100	5	60
13	20	100	50	105
14	20	100	100	155
15	20	250	5	60
16	20	250	50	105
17	20	250	100	155
18	60	25	5	60
19	60	25	50	105
20	60	25	100	155
21	60	100	5	60
22	60	100	50	105
23	60	100	100	155
24	60	250	5	60
25	60	250	50	105
26	60	250	100	155

The design matrix (Table 1) reflects four input parameters that will influence the composition of the resulting networks:

- ***RelativeChange*** - Percentage relative change of color when "chatted."

- ***InfluenceRng*** - Maximum distance of chat.

- ***FriendThresh*** - Agents within this range are considered "linked."

- ***EnemyThresh*** – Dependent variable; is calculated as FriendThresh plus 55, in order to preserve the same Friend to Enemy Distance (equivalent to the "neutral" range) as was present in the base scenario.

The CC scenario can be considered as a metaphor for a group of people establishing relationships based on shared interests or desires (color space proximity) and physical proximity (relative agent location). Agents are drawn toward agents with similar color and move away from agents of disimilar color. The closer agents are in location, the more frequently they "chat" each other, and thus, the closer they grow in color space. Eventually, cliques of "like-interest" agent form and are impacted by other agents and cliques. The input parameters varied in the design matrix affect these behavioral processes in straightforward ways.

2.3 Visualizing the Dynamic Network State

Part of a toolset to examine social network dynamics is the ability to analyze the ongoing agent interactions, behaviors, and network responses. Co-visualizing the various aspects (layers) of network dynamics can potentially provide powerful insight into the network.

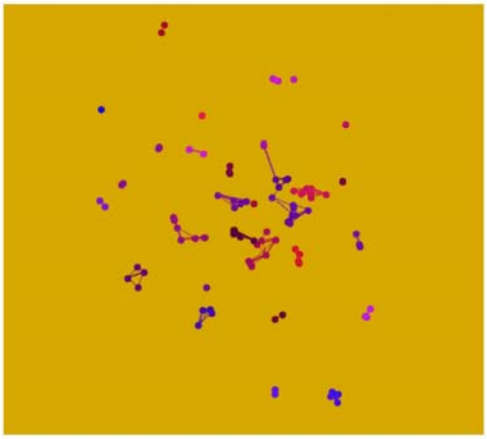


Figure 1. CC Scenario – Spatial View

The research team has done initial examination of the CC scenario using several visualization capabilites. Figure 1 is the spatial view provided by Pythagoras.

Figure 1 shows the agents at a time-step midway in the scenario. "Chats" are shown as lines between agents. This view, though, focuses on the location of the agent spatially.

Figure 2 shows four time-steps of an "inclination"-space view. In this image the location of the agents is based on their location in color space. The "redness" (0-255) of the agent is represented on the x axis. The "blueness" (0-255) of the agent is represented on the y axis. As the scenario proceeds left to right, top to bottom, note the congregation of agents into color groups. These groups do not represent the cliques formed though, because the spatial aspect is not represented.

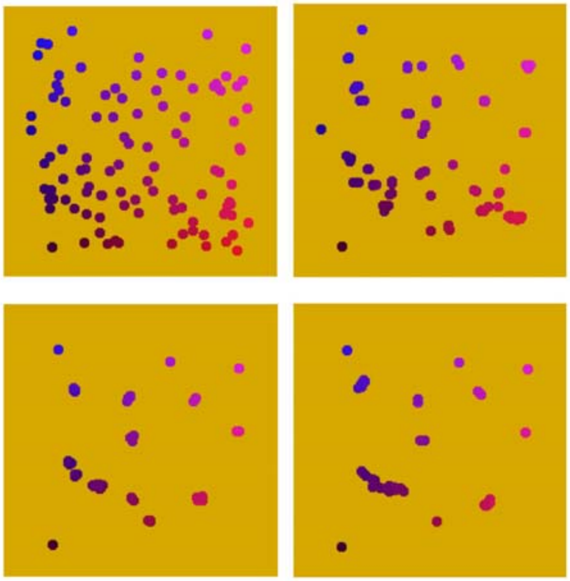


Figure 2. CC Scenario – Inclination Space View

FIgures 3 and 4 represent the same agent network , derived from the CC scenario, using the social network analysis "layout" generated by the R SNA plug-in and SoNIA software packages.

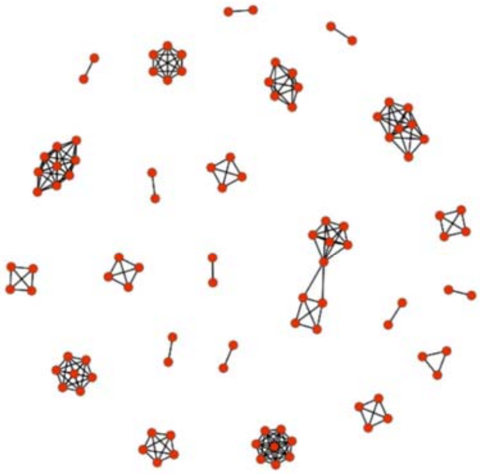


Figure 3. CC Scenario – Static Graph View

Figure 3 shows a static network layout representation of one of the CC time-steps using the default SNA layout algorithm. The SNA R package plots each time-step independently, not accounting for the layout defined in the previous time-step. The layout of each time-step is independent and as a result, the dynamic evolution is difficult to examine.

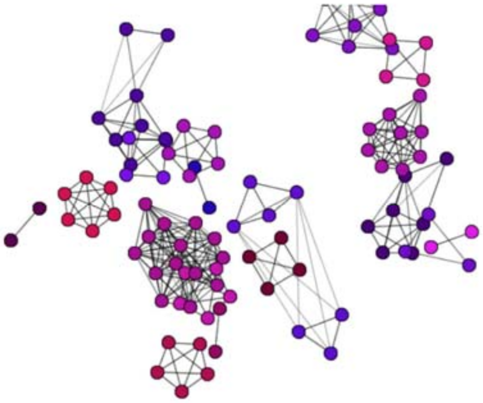


Figure 4. CC Scenario – Dynamic Graph View

Figure 4 shows a single time-step using the SoNIA application. SoNIA is designed to support dynamic time-series network data. As a result, the layout of any timestep can be based on the previous time step as a starting point. The result is a layout which displays the evolution of the network, but that can result in layouts that are not easily viewed statically.

It should be noted that Figures 2, 3 and 4 do not represent the spatial data shown in Figure 1 in any way... the "physical" location is ignored in these representations. In Figure 2 location represents color, and in Figures 3 and 4 the location is purely a function of the layout algorithm, which is designed to display the network in an uncluttered and easily-viewed manner, not the spatial location of the agents.

2.4 Social Network Analysis (SNA)

One of the team's goals is to begin to understand the utility of various SNA statistics in understanding the scenario dynamics and the result of data farming. Step one in this process was to delineate what outputs and analysis methods provide insight into network evolution and impact on agent behaviors.

SNA statistics fall into two classes: node statistics and network statistics. Node statistics include: betweenness, closeness, eigenvector centrality, and degree. Network statistics include: number of components, number of cliques, and average path length.

The study team decided to focus on node statistics initially and produced time-series output for every node of betweenness, eigenvector centrality and degree. Although data for 27 excursions of data farming was collected, it was decided to do an intial comparison of three excursions, where the primary variation was the color distance that defined what is considered a friend (a homophilic link). Excursions 0, 1, and 2 were examined.

Figure 5 represents one replication each from excusions 0, 1, and 2 as delineated in Table 1. The three plots represent the degree of each agent over time. The vertical axis is degree (the number of links associated with a node), the horizontal axis is time, and the axis going into the page is agent number. Figure 5 was generated using the PlotGL plugin to R.

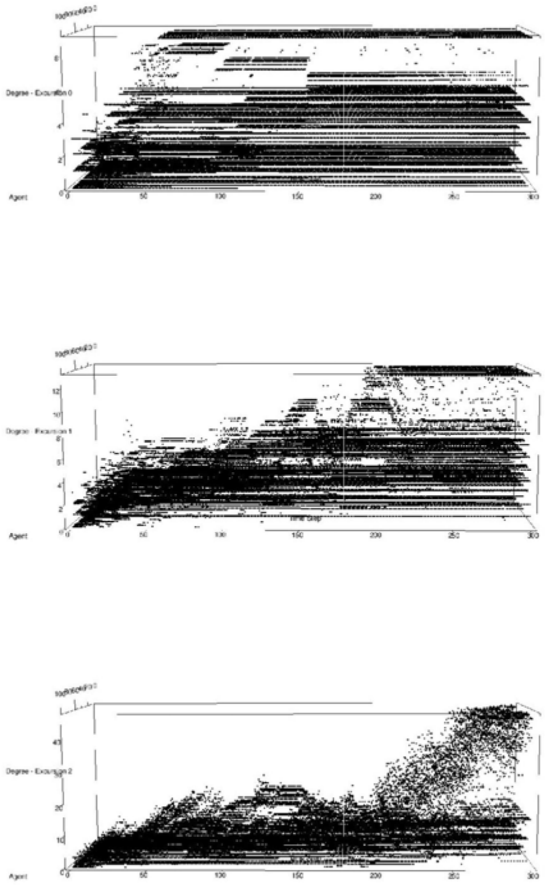


Figure 5. Centrality for Excursions 1-3

In Figure 5, various pattern differences, related to the evolution and devolution of cliques and components, can be discerned There are obvious differences between the excusions, with 0 and 1 appearing to reach covergence, but 2 never converging. It can be seen that some agents reach a steady-state and maintain it for some time, while other groups of agents particpate in behaviours which lead to the growth and reduction of degree for groups of agents.

2.5 Results

Two counter-intuitive results presented themselves. Excursion 2, in Figure 5c, shows that an increase in FriendThresh, that is, expanding the range and number of agents that an agent has homophilic links with in color space leads to increased instability in terms of clique formation. The initial assumption was that this would affect the size of the cliques and number of components. The unexpected result is that this increase prevents the stabilization of cliques and network components. Rather, it appears that this increase results in groups being able to "steal" members from other groups more easily.

Another interesting behavior is the Excursion 0 (Figure 5a) degree variation that occurs before equilibrium. In this case it appears that larger components are formed intially, but that they devolve into smaller groups over time. The team intends to investigate the set of replicates associated with this excursion to determine whether this behavior is consistent for this level of FriendThresh.

3.0 SUMMARY AND WAY AHEAD

Significant insight was gained by team members in delineating capabilities needed in a toolkit for the extraction and analysis of dynamic social data from models. The following capabilities will be needed for ongoing data farming research of basic social networks:

- **Synching of Visualization**: Various representations of the dynamic network are useful, but examining multiple views of the network time-step synced would provide powerful relational insights.
- **Equilibrium Time**: Determining whether equilibrium occurs and how long it takes is often the first step in analysis.
- **Data Farming Time Window Reduction Size**: Dynamic network analysis requires defining what constitutes a link, for example, a single interaction or multiple interactions over some time window. Being able to data farm this time window would provide analysts insight into network basics.
- **Node Statistic Capability**: Degree, betweenness, eigenvector, and

closeness need to be extractable for each node, time-step, replicate, and excursion and then represented effectively.

- **Network/Component Statistic Capability**: The number of cliques, and components, density, and others need to be acquired for each time step, replicate and excursion.

- **Newcomer/Leaving Effects**: Measure the effects of dynamic birth and death of agents.

- **Network Boundary Effects**: Data farm the impact of varying the size and extent of the network.

- **MOEs** (end-of-run vs. time-series) Both end-of-run and ongoing behaviors may be important.

The study team intends to continue to delineate tool capabilities for data farming social network models. We intend to accomplish the folowing tasks in the upcoming months:

- Document tools and methods identified in previous work.

- Define model output requirements for SNA analysis.

- Expand the toolkit to include additional network, node, and link statistics.

- Expand data farming methods for other network layers including weapon and resource interaction, spatial, communication, and multiple "inclination" parameters.

- Continue detailed analysis of CliqueCreator data farming results.

- Test use of tools and methods on other models (MANA, Netlogo scenarios).

- Begin delineating insurgent IED network scenario.

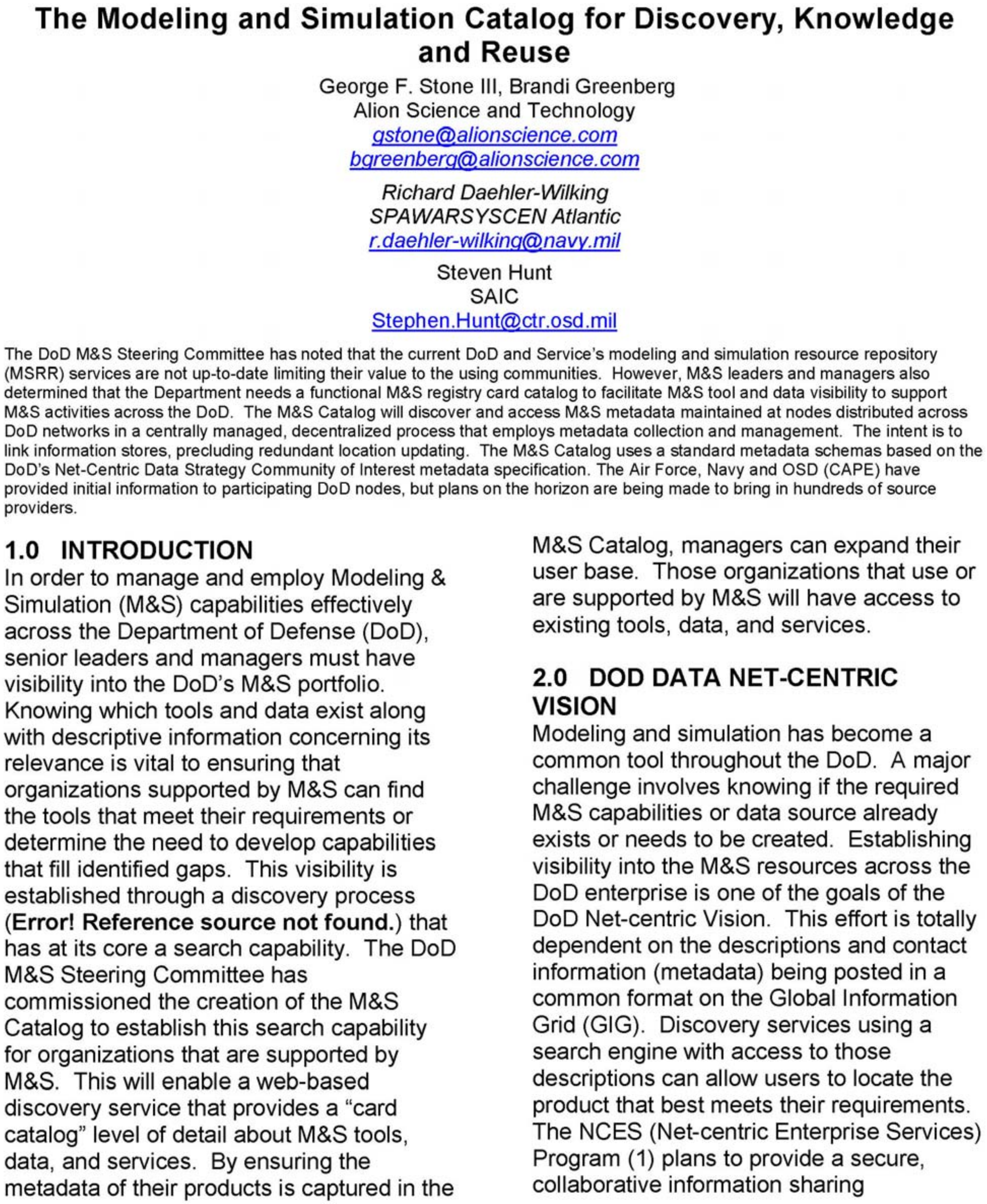
4.0 ACKNOWLEDGMENT

The authors would like to thank the members of Team 6 at International Data Farming Workshop 19 in Auckland, New Zealand and IDFW 20 in Monterey, California for their contributions, insights, and support.

5.0 REFERENCES

- Carrington, Peter J., Scott, John, Wasserman, Stanley, 2005, *Models and Methods in Social Network Analysis (Structural Analysis in the Social Sciences)*, Cambridge University Press
- Henscheid, Z., Middleton, D., and Bitinas, E. 2007. Pythagoras: An Agent-Based Simulation Environment, Scythe Issue 1: 40-44. Monterey, CA.
- Hoffman, F. and Horne, G. 1998. *Maneuver Warfare Science 1998*. United States Marine Corps Project Albert. Quantico, VA.
- Horne, G. 1997. Data Farming: A Meta-Technique for Research in the 21st Century, briefing presented at the Naval War College. Newport, RI.
- Horne, G. 1999. Maneuver Warfare Distillations: Essence Not Verisimilitude. Proceedings of the 1999 Winter Simulation Conference, eds. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 1147-1151. Phoenix, AZ.
- Horne, G. and Meyer, T. 2004. Data Farming: Discovering Surprise. Proceedings of the 2004 Winter Simulation Conference, eds. R. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 171-180. Washington, DC.
- Horne, G. and Meyer, T. 2005. Data Farming Architecture. Proceedings of the 2005 Winter Simulation Conference, eds. M. E. Kuhl, N. M. Steiger, F.B. Armstrong, and J. A. Joines, 1082-1087. Orlando, FL.
- Horne, G. and Meyer, T. January 2010. Scythe, Proceedings and Bulletin of the International Data Farming Community, Issue 7, Workshop 19, SEED Center for Data Farming, Monterey, CA.
- Horne, G. and Meyer, T. August 2010. Scythe, Proceedings and Bulletin of the International Data Farming Community, Issue 8, Workshop 20, SEED Center for Data Farming, Monterey, CA.
- Kleijnen, J., Sanchez, S., Lucas, T., and Cioppa, T. 2005, A User's Guide to the Brave New World of Designing Simulation Experiments, INFORMS Journal on Computing, 17(3): 263-289. Hanover, MD.
- PlotGL R Package (http://cran.r-project.org/web/packages/plotgl/index.html)
- SNA R Package (http://cran.r-project.org/web/packages/sna/index.html)
- SoNIA Social Network Image Animator (http://www.stanford.edu/group/sonia/index.html)
- Wasserman, Stanley, Faust, Katherine, 1994, *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*, Cambridge University Press

1.9 The Modeling and Simulation Catalog for Discovery, Knowledge and Reuse



The Modeling and Simulation Catalog for Discovery, Knowledge and Reuse

George F. Stone III, Brandi Greenberg
Alion Science and Technology
gstone@alionscience.com
bgreenberg@alionscience.com

Richard Daehler-Wilking
SPAWARSYSCEN Atlantic
r.daehler-wilking@navy.mil

Steven Hunt
SAIC
Stephen.Hunt@ctr.osd.mil

The DoD M&S Steering Committee has noted that the current DoD and Service's modeling and simulation resource repository (MSRR) services are not up-to-date limiting their value to the using communities. However, M&S leaders and managers also determined that the Department needs a functional M&S registry card catalog to facilitate M&S tool and data visibility to support M&S activities across the DoD. The M&S Catalog will discover and access M&S metadata maintained at nodes distributed across DoD networks in a centrally managed, decentralized process that employs metadata collection and management. The intent is to link information stores, precluding redundant location updating. The M&S Catalog uses a standard metadata schemas based on the DoD's Net-Centric Data Strategy Community of Interest metadata specification. The Air Force, Navy and OSD (CAPE) have provided initial information to participating DoD nodes, but plans on the horizon are being made to bring in hundreds of source providers.

1.0 INTRODUCTION

In order to manage and employ Modeling & Simulation (M&S) capabilities effectively across the Department of Defense (DoD), senior leaders and managers must have visibility into the DoD's M&S portfolio. Knowing which tools and data exist along with descriptive information concerning its relevance is vital to ensuring that organizations supported by M&S can find the tools that meet their requirements or determine the need to develop capabilities that fill identified gaps. This visibility is established through a discovery process (**Error! Reference source not found.**) that has at its core a search capability. The DoD M&S Steering Committee has commissioned the creation of the M&S Catalog to establish this search capability for organizations that are supported by M&S. This will enable a web-based discovery service that provides a "card catalog" level of detail about M&S tools, data, and services. By ensuring the metadata of their products is captured in the M&S Catalog, managers can expand their user base. Those organizations that use or are supported by M&S will have access to existing tools, data, and services.

2.0 DOD DATA NET-CENTRIC VISION

Modeling and simulation has become a common tool throughout the DoD. A major challenge involves knowing if the required M&S capabilities or data source already exists or needs to be created. Establishing visibility into the M&S resources across the DoD enterprise is one of the goals of the DoD Net-centric Vision. This effort is totally dependent on the descriptions and contact information (metadata) being posted in a common format on the Global Information Grid (GIG). Discovery services using a search engine with access to those descriptions can allow users to locate the product that best meets their requirements. The NCES (Net-centric Enterprise Services) Program (1) plans to provide a secure, collaborative information sharing environment with access to decision-quality information through a Service Oriented Architecture (SOA) (2) that enables achievement of the DoD's data strategy. An NCES goal is to increase mission effectiveness by enhancing process execution across the DoD. The NCES Program provides four product lines:

1. Enterprise Service-Oriented Architecture Foundation
2. DoD Enterprise Collaboration
3. Content Discovery and Delivery
4. Web Portal

3.0 M&S CATALOG DISCOVERY PROCESS

The main product that will be focused on in this paper is "discovery"--the ability to locate data assets through a consistent and flexible search. The discovery process starts when an organization or developer begins to generate a new M&S product or capability, and it is enabled by the creation of metadata about that product or capability. The process requires that metadata be in an electronic format and accessible to some type of search tool or mechanism through which potential users can find the metadata and access the product or service. The steps associated with the M&S Catalog Discovery Process are depicted in Figure 1 below.

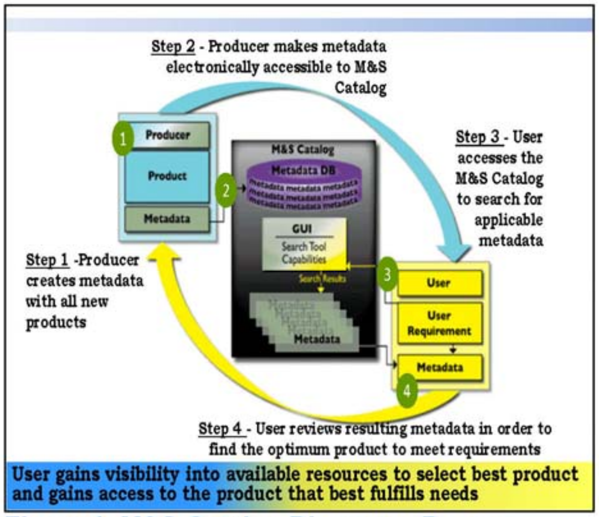

Step 2 - Producer makes metadata electronically accessible to M&S Catalog
1
Producer
Product
Metadata
2
M&S Catalog
Metadata DB
metadata metadata metadata
GUI
Search Tool Capabilities
Search Results
Metadata
Step 3 - User accesses the M&S Catalog to search for applicable metadata
3
User
User Requirement
Metadata
4
Step 1 -Producer creates metadata with all new products
Step 4 - User reviews resulting metadata in order to find the optimum product to meet requirements
User gains visibility into available resources to select best product and gains access to the product that best fulfills needs

Figure 1. M&S Catalog Discovery Proces

4.0 METADATA

One facet of an enterprise-wide discovery capability is the ability to consistently describe data assets. The description of data is called metadata; defined as being "data about data." A common specification for the description (metadata) of data assets allows for a comprehensive capability that can locate all data assets across the enterprise regardless of format, type, location, or classification. To facilitate data asset discovery, the DoD has developed the DoD Discovery Metadata Specification (DDMS) (4) as the common set of descriptive metadata elements to be associated with each data asset visible to the enterprise discovery capability. The DDMS defines discovery metadata elements for resources posted to community and organizational shared spaces. The DDMS specifies a set of information fields that are to be used to describe any data or service asset that is made known to the Enterprise, and it serves as a reference for developers, architects, and engineers by laying a foundation for Discovery Services. The DDMS will be employed consistently across the Department's disciplines, domains and data formats.

5.0 M&S COMMUNITY OF INTEREST DISCOVERY METADATA SPECIFICATION (MSC-DMS)

An element of the DoD Net-Centric Data Strategy (3) is the formation of communities of interest (COI) to address data exchange issues common to that community. One of the tasks of a COI is to establish a common specification for the discovery metadata to be used within that community. The community discovery metadata specifications (DMS) are to use the DDMS as a foundational specification and add those metadata elements that are required for the community to accurately describe their products. It is critical that the community specifications include metadata elements that enable product owners to express the difference between their products and other similar products. While the common metadata elements do not have to be identical to the DDMS, there must be a mapping from the common metadata elements in the community DMS to the DDMS.

The M&S COI (MSC) Discovery Metadata Specification (MSC-DMS) (5) specifies the set of information fields that are to be used to describe M&S tools, data or services which are to be made known to the Enterprise. It serves as a reference for metadata associated product developers, architects, and engineers. The DDMS and other standards, practices, and approaches have been cross integrated to formulate the MSC-DMS to be used across the Communities and Services for tagging M&S assets that will be made accessible via the GIG. All activities that publish the availability of M&S assets should use the MSC-DMS.

6.0 MSC-DMS M&S CATALOG ASSOCIATION

The early design requirements of the M&S Catalog recognized the importance of aligning the metadata used in the Catalog with the standards being established in the DoD enterprise and specifically within the M&S Community. Since the M&S Catalog is the primary major project currently utilizing the MSC-DMS, there has been close coordination between the teams developing each. Early in the first year of the M&S Catalog project, the develop team, to include representatives from the Navy's Space and Naval Warfare (SPAWAR) Systems Center Atlantic, the Air Force Agency for M&S (AFAMS), CAPE JDS, and the DoD M&S Coordination Office (M&S CO), reviewed and improved on a mapping from the individual Service MSRRs to the M&S Community of Interest (COI) Discovery Metadata Specification (MSC-DMS). Many suggestions for improvements to the MSC-DMS were developed and later incorporated into version 1.2 of the MSC-DMS. There has been ongoing coordination between the two teams to ensure that the development of the M&S Catalog and MSC-DMS remain aligned.

As part of their efforts to address M&S capability gaps common to organizations across the DoD enterprise, the DoD M&S Steering Committee commissioned an M&S Catalog search tool to provide a web-based discovery service focused on M&S tools, data, and services.–The search tool capabilities selected for the M&S Catalog were guided by interviews with senior leaders, users, and technical personnel in the communities that participate in the M&S Steering Committee. The intent is to make the search tool as intuitive and effective as possible, to guide a user quickly to a manageable set of alternatives to evaluate. Additionally, in response to the request of senior level managers, the tool will have the capability to perform analysis of the characteristics of the search result set of resources.

The resulting visibility into the M&S world will provide significant benefits throughout DoD. Resource owners can use the catalog to maintain their own inventories as well as identify new customers. Resource seekers can rapidly find what they need and identify potential cost avoidances by learning of existing efforts. The department will achieve better resource management by ensuring resources are not applied to create existing capabilities, but instead focus on those areas where capabilities are lacking.

7.0 SOURCES

The key to the value of the M&S Catalog is the breadth and accuracy of the information it contains. A significant effort is being under taken to encourage organizations across the DoD enterprise to integrate the information about their products with the M&S Catalog. Metadata can be accepted from a collection such as a service M&S Resource Repository (MSRR) or directly from the manger of a product. The vision is to interface as closely to the origin of the metadata as possible so that the motivation to keep it current is high.

One of the primary tasks in the next phase of the M&S Catalog development is to significantly increase the number of sources integrated with the Catalog. In order to enable this, a major outreach program has been initiated and tools will be developed to reduce the level of effort required to create, maintain and integrate metadata. M&S Catalog products include any resource that can be used to support an M&S effort:

1. Services – Organizations that can provide design, development, or analysis support.
2. Tools – Software and hardware to support models and simulations.
3. Data – Data the model or simulation requires.
4. Subject Matter Experts – Domain experts that can provide guidance on the selection of model parameters, problem specific data, and/or validation for models.

8.0 HOW THE M&S CATALOG FITS INTO THE DISCOVERY PROCESS

The discovery metadata search mechanism to interface between the producers and consumers is the M&S Catalog. The metadata that is accessible through the M&S Catalog and the functionality of the discovery tool will determine how well a user can find the product that best meets their needs. The metadata format within the M&S Catalog needs to contain the elements that the user's community uses to differentiate the products they use. The user interface and the flexibility of the search tools will have a large impact on how successful the users are connected to the optimum products for their requirements. The design, format and content of all other elements of the discovery process must integrate smoothly into the M&S Catalog in order for the process to function well.

While many organizations have expressed an interest in providing metadata to the M&S Catalog, often there are limited resources with which to produce and transform metadata. It has become apparent that the level of effort placed on the source organizations must be as minimal as possible to enable their participation. Providing tools and processes to aid them will not only increase the likelihood of the metadata integration, it will also support consistency in the metadata content and format.

9.0 CAPABILITIES

Based on the lessons learned from the earlier phases and the requirements generated from interviews with representatives of the communities belonging to the M&S Steering Committee, it was decided that the third phase of the M&S Catalog project would migrate to a COTS tool that offered a good fit to the desired capabilities. Market research was conducted, the offerings of several vendors were compared, and finally the Endeca Information Access Platform was selected and acquired (see figure 2). The current phase of the M&S Catalog now offers:

- Facetted search –Dynamically guided navigational search offering selections based on community driven taxonomies. Each subsequent selection searches within the previous results. Previously selections can be removed individually. This allows the user to create their own taxonomy through the metadata elements they select.
- Flexible support of different source metadata structures, including unstructured documents.
- "Tag clouds" – term / phrase occurrence analysis within the results set.
- Support of quantitative analysis on the search results(e.g., how many tools deal with air-to-air by source organization)
- Keywords – traditional search for specific strings that can be applied at any point in the navigational search process
- User determined search result format – the user selects the metadata elements to be displayed in the search results

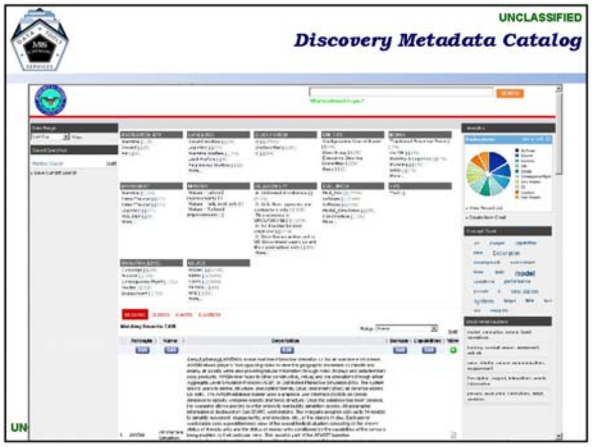


Figure 2. M&S Catalog User Interface

10.0 FUTURE EFFORTS OF THE M&S CATALOG PROJECT

In the coming months, efforts in the M&S Catalog project will be aimed at improving the capabilities of the tool itself, improving the data model used by the M&S Catalog, continuing outreach to sources (both new and current), development of metadata creation and transformation tools to enable source organization metadata efforts and search federation with other search engines.

Improvements of the tool itself will include:

- Flexible support of DoD-relevant taxonomies.
- Multiple user-interface screens addressing different needs.
- Resource ranking & comments by users
- Forums

DoD-relevant taxonomies were important in previous search tools because of the rigid structure of the search capability. The facetted search capability (presented in **Error! Reference source not found.** and **Error! Reference source not found.**) allows the user to develop their own path to the resources that meet their requirements. In essence it allows the user to develop their own tailored taxonomy that is relevant to the particular resource requirements that guide their search. The different taxonomies found in different communities are driven by different search criteria or different descriptions. As the organizations providing metadata and user search criteria increase, the number of facetted search categories may grow significantly. Placing all supported facetted search selections on one screen may reduce the usability of the M&S Catalog. Different user interface pages will be developed with subsets of the facetted search options. The community taxonomies will be used to guide the determination of what facetted search selections will be listed on each user interface page.

In the search for pedigree, the best input often is the experience of others who have used a resource. The M&S Catalog will be adding the capability of users to rank and comment on resources. Additionally, often experience users can be a great resource for new users to determine the best methods with which to access or use a particular tool, data source or service. In order to take advantage of sharing of ideas and experience, a forum capability will be added to the M&S Catalog.

The number one priority of the M&S Catalog for 2010 is increasing participation by sources. Our efforts include outreach to the sources themselves, recruitment of senior leadership of the DoD communities enabled by M&S, and assistance to sources that want to participate. Such assistance includes development of tools to assist in metadata creation and maintenance, mapping to the M&S Catalog data model and electronic interface with the M&S Catalog.

11.0 CONCLUSION

Finally, upcoming work in 2010 includes the federation of the M&S Catalog with other search engines. The DoD Data Net-centric Vision established the DDMS as the common discovery metadata for federated searches. The M&S Catalog metadata must be exportable in a DDMS identifiable format. Additionally, the M&S Catalog must be capable of accepting DDMS formatted search queries. These capabilities will initially be targeting federation with the DISA Enterprise Catalog.

The M&S Catalog is available to anyone with a DoD-approved Common Access Card (CAC) or External Certificate Authority (ECA) security certificate at https://MSCatalog.osd.mil.

12.0 REFERENCES

(1) Net-Centric Enterprise Services (NCES) User Guide, Version 1.2, 12 March 2008, http://www.disa.mil/nces/users_guide_03122008.pdf

(2) Reference Architecture Foundation for Servicer Oriented Architecture, Version 1,0, Committee Draft 02, 14 October 2009, OASIS, http://docs.oasis-open.org/soa-rm/soa-ra/v1.0/soa-ra-cd-02.pdf

(3) Department of Defense Net-Centric Data Strategy, DoD CIO Memorandum, May 9, 2003

(4) Department of Defense Discovery Metadata Specification (DDMS) – Version 2.0 – 16 July 2008, Deputy Assistant Secretary of Defense (Deputy Chief Information Officer).

(5) Modeling and Simulation (M&S) Community of Interest (COI) Discovery Metadata Specification (MSC-DMS) – Version 1.2, February 20, 2009.

13.0 ACKNOWLEDGMENT

The authors wish to thank the Department of Defense Modeling and Simulation Senior Leadership Council for their continuous interest and support of the M&S Catalog.

1.10 Leveraging M&S software to build Marine Survival Craft Training Simulators

MODSIM WORLD Conference & Expo

October 13–15, 2010
Hampton, Virginia

Leveraging M&S software to build Marine Survival Craft Training Simulators

Sebastien Loze

PRESAGIS   VMT virtual marine technology

October 15, 2010

MODSIM WORLD Conference & Expo

Statistics

- **16%** - Percentage of total maritime lives lost over a 10 year period that were attributed to lifeboat accidents
- **12.5%** - Percentage of lifeboat drills that were performed unsatisfactorily during a 2009 inspection campaign
- **1.4%** - Percentage of ship inspections that identified lifeboat deficiencies serous enough to warrant detention

"Anyone using a lifeboat, be it in a drill or genuine evacuation, runs a risk of being injured or even killed."
UK Maritime Accident Investigation Branch – 2001

MARINE REPORTS - 2002 - M02L0061
Transportation Safety Board of Canada (TSB) Marine Investigation Report
Crew Member Lost Overboard

Bulk Carrier Keat
Near Verchères, Quebec
16 July 2002
Report Number M02L0061

Summary
On 16 July 2002, in fair weather and good v
Quebec, under the conduct of a pilot. In the
port lifeboat ready for sea. The bosun was
suddenly released and struck him on its de
The alarm was raised and several life rings

At about 1055 local time on Saturday 20 Janu
and flag in Seaforth Dock, Liverpool, the port
from the davits into the basin waters. Three o
various injuries, one serious head injury.

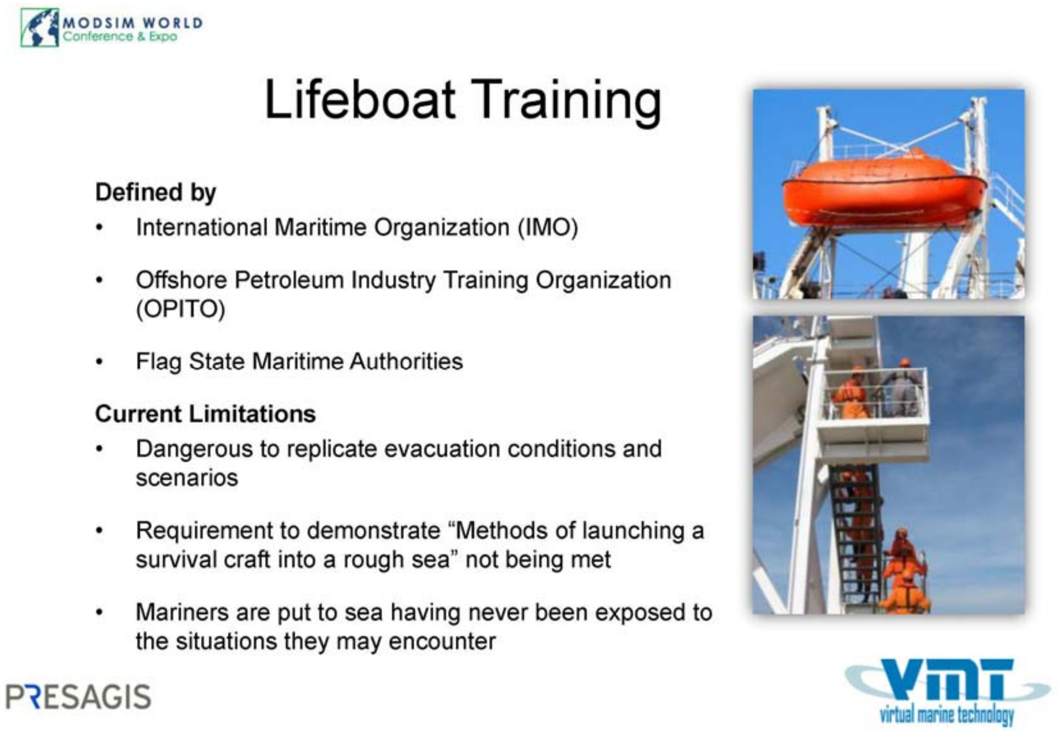
Lifeboat Accident Kills One; Injures Two More

Portland Press Herald   January 14, 2004

URL:
http://www.rigzone.com/news/article.asp?a_id=10424

One man was killed and two co-workers were injured Tuesday when a

PRESAGIS   VMT virtual marine technology



MODSIM WORLD Conference & Expo

Lifeboat Training

Defined by

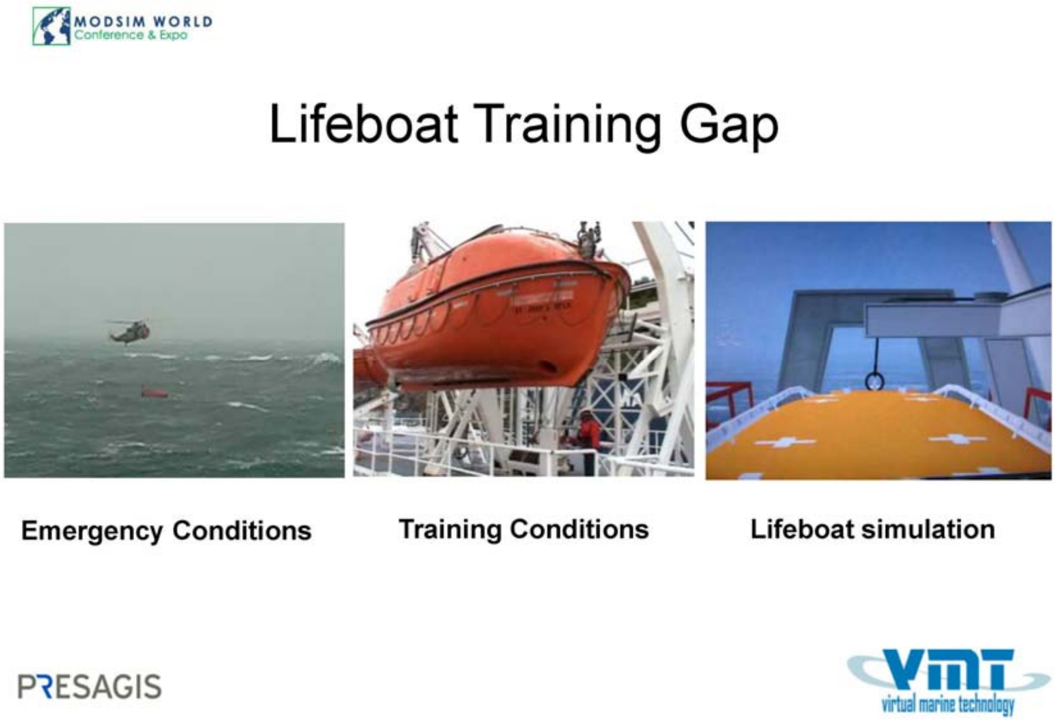
- International Maritime Organization (IMO)
- Offshore Petroleum Industry Training Organization (OPITO)
- Flag State Maritime Authorities

Current Limitations

- Dangerous to replicate evacuation conditions and scenarios
- Requirement to demonstrate "Methods of launching a survival craft into a rough sea" not being met
- Mariners are put to sea having never been exposed to the situations they may encounter

PRESAGIS

VMT virtual marine technology



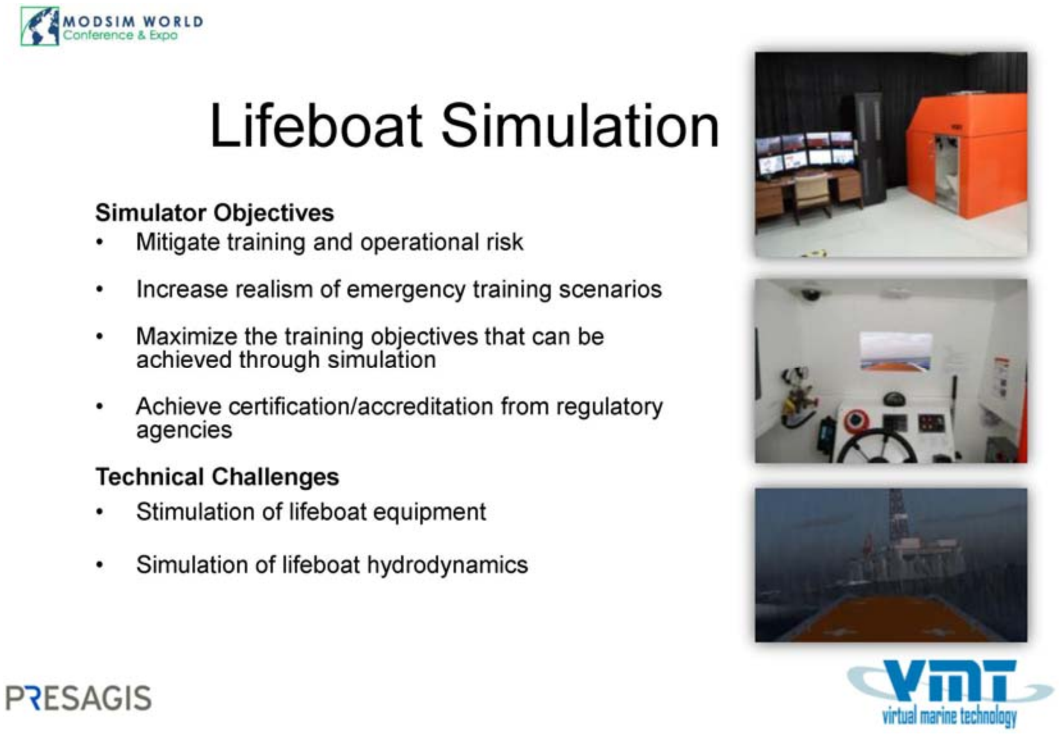
MODSIM WORLD Conference & Expo

Lifeboat Training Gap

Emergency Conditions **Training Conditions** **Lifeboat simulation**

PRESAGIS

VMT virtual marine technology



MODSIM WORLD Conference & Expo

Lifeboat Simulation

Simulator Objectives

- Mitigate training and operational risk
- Increase realism of emergency training scenarios
- Maximize the training objectives that can be achieved through simulation
- Achieve certification/accreditation from regulatory agencies

Technical Challenges

- Stimulation of lifeboat equipment
- Simulation of lifeboat hydrodynamics

PRESAGIS

VMT virtual marine technology

MODSIM WORLD Conference & Expo

VEGA PRIME Marine

PRESAGIS

VMT virtual marine technology

Open Programming Architecture

- Create applications using supplied wave models or use your own custom wave models
- Incorporate completely foreign wave simulation algorithms and have them incorporated automatically into the provided rendering environment
- Apply the same calculations to the visual and non-visual (i.e. host computer calculating ships motions, forces, and dynamics)
- Produce complex wave models using an open and intuitive interface

Multiple Ocean Types

- Position to any view point
  - Fixed location
  - Observer-centered
  - Surf zone
  - Large Area / Round Earth

Synthesized Surfaces

- Physically correct wave model out of the box
- Maritime effects
- Customizable pre-defined ship motion strategies
- Short and Long crested waves
- Environmental and local reflections

Synthesized Surfaces

- 13 sea states described by the Beaufort scale
- 9 sea states described by the Spectral Ocean Wave Model

- Multiple user-defined ocean definition parameters

MODSIM WORLD Conference & Expo

Surf Zone

- Shallow water modeling and coastline effects
  - Breaking waves
  - Cusp Surf
  - Sandbars
  - Depth and shoreline transitions
  - Wave effects on vehicle motion
  - Correct wave behavior
  - Seamless Transition from shallow to deep water

PRESAGIS

VMT virtual marine technology

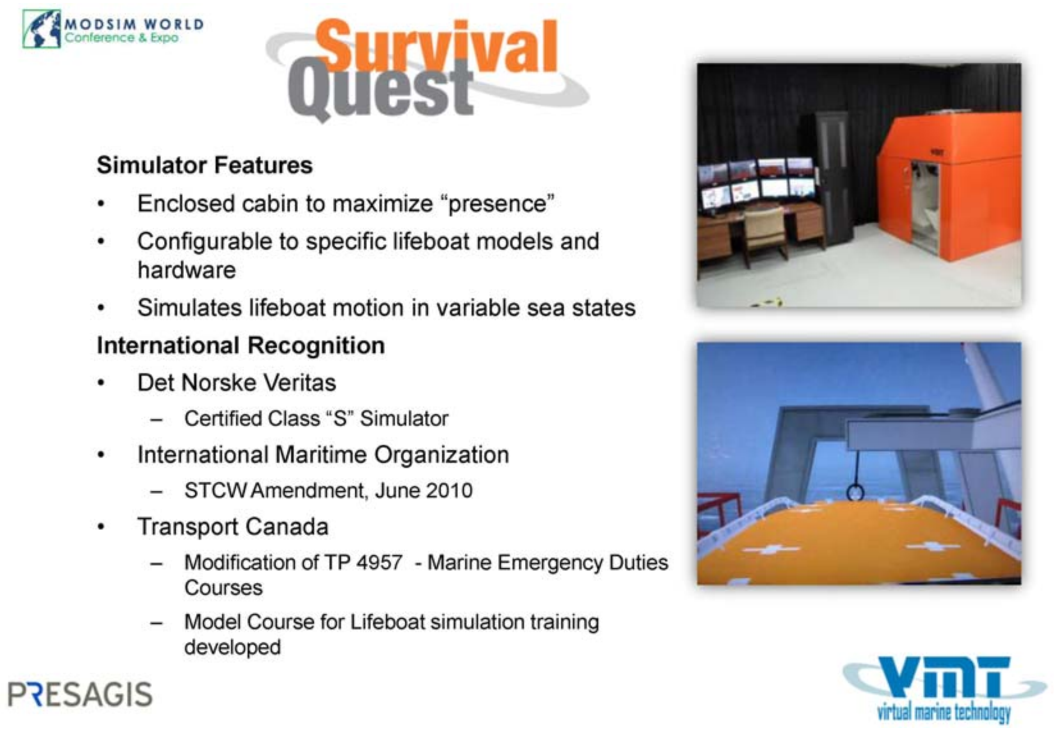
MODSIM WORLD Conference & Expo

Marine Effects

- User-defined vessel characteristics:
  - Bow waves
  - Stern
  - Hull
  - Size and shape correspond to the size, shape, and speed of the vessel
  - Interaction with the ambient water waves
  - Visual aid in determining the speed, maneuvering, and turning of the vessel
- Customizable ship motion strategy for correct behavior of objects / vessels on the ocean

PRESAGIS

VMT virtual marine technology



MODSIM WORLD Conference & Expo

Survival Quest

Simulator Features

- Enclosed cabin to maximize "presence"
- Configurable to specific lifeboat models and hardware
- Simulates lifeboat motion in variable sea states

International Recognition

- Det Norske Veritas
  - Certified Class "S" Simulator
- International Maritime Organization
  - STCW Amendment, June 2010
- Transport Canada
  - Modification of TP 4957 - Marine Emergency Duties Courses
  - Model Course for Lifeboat simulation training developed

PRESAGIS

VMT virtual marine technology

MODSIM WORLD Conference & Expo

Looking Ahead

High Speed Boats

- Training for Coast Guard, Navy and Waterborne Law Enforcement
- Vessel planing at 40+ knots
- Launch and recovery in chaotic wave environments

PRESAGIS

VMT virtual marine technology

Thank You

Virtual Marine Technology
tyler.brand@vmtechnology.ca

Presagis
Sebastien Loze
Product Marketing Manager – Visualization & Simulation
Sebastien.Loze@presagis.com

2.0 ENGINEERING AND SCIENCE TRACK

2.1 Executable Architecture Research at Old Dominion University

Executable Architecture Research at Old Dominion University

Andreas Tolk
Old Dominion University
atolk@odu.edu
Johnny J. Garcia
Old Dominion University (SimIS Inc.)
johnny.garcia@simisinc.com
Edwin A. Shuman
Old Dominion University (MITRE)
ashuman@mitre.org

Abstract. Executable Architectures allow the evaluation of system architectures not only regarding their static, but also their dynamic behavior. However, the systems engineering community do not agree on a common formal specification of executable architectures. To close this gap and identify necessary elements of an executable architecture, a modeling language, and a modeling formalism is topic of ongoing PhD research. In addition, systems are generally defined and applied in an operational context to provide capabilities and enable missions. To maximize the benefits of executable architectures, a second PhD effort introduces the idea of creating an executable context in addition to the executable architecture. The results move the validation of architectures from the current information domain into the knowledge domain and improve the reliability of such validation efforts. The paper presents research and results of both doctoral research efforts and puts them into a common context of state-of-the-art of systems engineering methods supporting more agility.

1.0 INTRODUCTION

This paper introduces two ongoing related PhD efforts at Old Dominion University. Both efforts contribute to the topic of Executable Architecture research. Being members of the active M&S work force in Hampton Roads, both PhD candidates collected valuable experiences in projects and research. Embedding these experiences into the scholastic education within the Modeling and Simulation program of Old Dominion University ensures academically valuable results that promise to be practically useful as well.

The mentor for this work has experiences in the academic and practical realm as well. In a study on Active Layered Theatre Ballistic Missile Defense (ALTBMD) for NATO, he was member of an international team that used the "Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance (C4ISR) Architecture Framework," which evolved later into the "Department of Defense (DoD) Architecture Framework (DoDAF)" and the "NATO Architecture Framework (NAF)," to define ALTBMD architecture and execute them using simulation systems like the German "Tactical Missile Defense Simulator (TMDSIM)," the US "Extended Air Defense Simulator (EADSIM)," and the US "Extended Air Defense Test Bed (EADTB)" to evaluate and compare the different architecture proposals [1].

Each PhD effort is an individual contribution, but presenting them together in this paper allows focusing on the synergy between the research results. Both contributions address important gaps in the body of knowledge for executable architecture research. To do this, the paper is structured as follows. Section two will deal with a state of the art overview and present a summary of related research. The third section will focus on the necessity for a more formal approach to executable architecture, comprising the definition of elements that are pivotal for such an architecture, evaluating alternative modeling languages to model what needs to be executed, and a formalism allowing to introduce the necessary rigor. The fourth section introduces the idea of executable contexts. While the executable architecture represents the system under development, the executable context represents the operational environment the system will be applied in. Finally, the concluding section will synthesize both efforts and place them into a broader research agenda.

2.0 STATE OF THE ART

The overview in this section is neither complete nor exclusive. However, it shows the trend of recent developments, in particular for defense related architecture evaluations. The general underlying idea motivating the use of executable architectures is to enable the evaluation of dynamic aspects. A system's architecture is the static blueprint of a system that identifies who (function) is doing what (capability) where (component). Executable architectures allow furthermore to evaluate when (time) something is done. Dead locks, internal loops, and other related problems can be detected in the definition phase of the system.

Zinn [2] investigated the utility of using DoDAF architecture products to provide needed data for agent based simulations. This was accomplished by means of a case study where architecture data from a proposed Air Operations Center architecture was used in the combat model System Effectiveness Analysis Simulation (SEAS). The research concluded that DoDAF, if implemented properly, does provide the needed information for developing agent-based simulations. Zinn proposed a process of taking information from DoDAF architectures and importing it into an agent-based simulation. To model process information, Zinn used information contained in the OV-5 and OV-6a (IDEF3) to feed the agent-based simulation. The OV-5 provides the process and information flow, while the OV-6a provides the decision logic associated with the process.

Wagenhals et al. [3] provide a description of an architecting process based on the object-oriented Unified Modeling Language (UML). They describe a mapping between the UML implementations and an executable model based on Colored Petri nets. They examine DoDAF product sufficiency in terms of the Colored Petri Nets (CPN) simulations end state objective. Wagenhals et al. focus on the UML Sequence Diagram (OV6c), the UML Collaboration Diagram (OV5b) and the Class Diagram (OV5a – with extensions).

In 2005, Ziegler and Mittal [4] described the translation of DoDAF compliant architectures into DEVS simulations. They provided a set of DoDAF foundational Views and related UML diagrams for construction of DEVS-based simulations

In 2006, Mittal [5] addressed the question of extending DoDAF to support integrated DEVS-based modeling. His work cited DoDAF's shortcomings, to include his assertion of ill-defined information exchanges, the need for a coupling of entities, activities, and nodes, and a need to identify ports associated with activity-to-activity communication (since DEVS is a port-based modeling construct). He defined two new OV products, the OV-8 and the OV-9, as extensions of the DoDAF. The OV-8 addresses activities and their logical interface information. The OV-9 maps nodes, entities, and activities. This is similar conceptually to Activities-based methodology [6]. Mittal asserted the need for the OV-8 and OV-9 as intermediate precursor products in the development of the DEVS simulation. Mittal used the OV-5 activity model, the OV-6c (Sequence Diagram) and the OV-6a (Rules diagram – IDEF3), as a basis for generating a DEVS-based simulation.

In 2006, Mittal [7] described a means for semantically strengthening the critical OV-6a Rules Model, through application of Domain Meaning, Units of Measure (UOM), and formatting to domain specific rules, thereby removing ambiguity and aiding in translation of static to dynamic architectures.

In 2009, Risco-Martin et al. [8] described the essential mappings between UML and DEVS modeling. That work focused on the UML Structure and Behavior models that contribute to the development of a DEVS-based system model. Those UML models are the Component Diagram, the State Machine, the Sequence Diagram, and the Timing Diagram.

3.0 A FORMAL APPROACH TO EXECUTABLE ARCHITECTURES

When evaluating the current approaches to derive executable architectures from static architectures, such as captured in DoDAF or comparable frameworks, it becomes obvious that the objective of these efforts is the use of dynamic simulation software to evaluate architecture models [5]. However, the current research is more concerned about concrete methods and tools, like the use of DEVS and DoDAF, the use of CPN and DoDAF, and similar projects.

The objective of the first PhD thesis is therefore to contribute to a theory of executable architectures. First results of this research are presented in [9]. The research derived from the observations of current approaches that it can be hypothesized that three categories are needed to define the necessary components for an executable architecture.

3.1 Elements

Elements define the static WHO, WHAT, and WHERE parts of an architecture. The elements provide the conceptual, structural, functional and state descriptions needed to describe and analyze a system. An architecture framework helps us to establish the boundaries for the discussion and to give it context and perspective. Examination of relevant elements of an architecture framework from conceptual, structural, functional and state perspectives helps to scope the topic of discussion.

3.2 Language

A modeling language allows us to instantiate the specifics of our architectural subset by describing both static and dynamic aspects of a system. The modeling language provides graphical, symbolic, standard notations designed to address various kinds of analysis and inquiry. A specific example of this would be a System Modeling Language (SysML) instantiation of the DoDAF OV-5, Operational Activity Diagram. That SysML diagram allows us to describe system behavior, or the functional system perspective.

3.3 Modeling Formalism

A modeling formalism for executable architectures should holistically describe the elements of an executable architecture using a standard mathematical notation. This ties the WHO, WHAT, WHERE, and WHEN together in a consistent and complete way. Traditionally, validation and verification supports this task. The formalism provides the mathematical frame to really prove that all functions are provided, interconnected, etc. The DEVS formalism is a promising first candidate. The elements of an executable architecture should be described using a modeling formalism, and minimally in the context of DEVS.

Figure 1 shows the concept triangle, including some examples for the components.

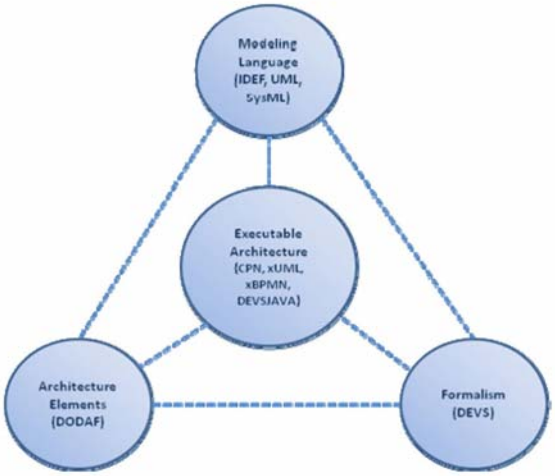

Modeling Language (IDEF, UML, SysML)
Executable Architecture (CPN, xUML, xBPMN, DEVSJAVA)
Architecture Elements (DODAF)
Formalism (DEVS)

Figure 1. Concept Triangle

A theory of executable architectures must ensure that the architecture can be described completely and consistently through all three components. All elements captured in the Architecture Elements need to be part of the formalism and should be the subject or object of activities modeled with the Modeling Language. The Modeling Language must be the subject of the formalism and should not use elements that are not captured in the Architecture Elements. The Formalism must bind elements and actions together and provide the mathematics to support validation and validity.

The first results published in [9] already show how to show alignment between Architecture Elements and Modeling Languages and apply metrics to the degree of alignment. Shuman showed, e.g., that for executable architectures the use of the SysML may be preferable to the use of the UML.

He also showed that the Fishwick modeling taxonomy [10], which distinguishes between conceptual, declarative, functional, constraint-oriented and spatial models, has significant value to support the three components of the concept triangle for executable architecture and provides foundational input for the general theory.

Such a theory will help to transfer valuable and practically relevant results between the various contributions so far. It will also support transferability of architecture artifacts, as de facto the components elements, language, and formalism must become a general meta-model of relevant approaches allowing to derive specialized solutions, like used in the examples described earlier.

4.0 ADDING EXECUTABLE CONTEXT

The focus so far has been on the system. Validating system architectures and assessing the contribution and efficiency of the specified systems before a system is built is the objective supported by the research of the second PhD effort. As pointed out before, the current state of the art of validation in practice is limited to static methods answering questions regarding who is doing what where. Executable architectures support system behavior analysis. They support examination of system timing questions (WHEN). They address questions related to the WHY and HOW of system behavior. In addition to this, executable architecture should address system context as well. Garcia presented the theory in [11] and showed an application example in [12].

Buede introduces the system's context as "a set of entities that can impact the system but cannot be impacted by the system. The entities in the systems context are responsible for some of the systems requirements." [13, p. 38] He also introduces external systems that interact with the system under development. Together, they introduce the system environment. Figure 2 captures these ideas.

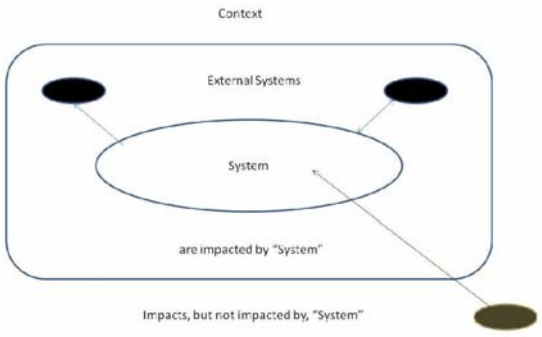


Figure 2. System, External Systems, and Context

While the executable architectures allow evaluation of system behavior (such as deadlocks and infinite loops), Fig. 2 shows that significant effects to system behavior will occur as a result of interaction with other external systems or even as a result of interactions between external systems. It is possible that the same category of dynamic problems that are evaluated in the previous section for the system's internal components by executable architectures – such as deadlocks between components – can occur between the system and external systems in the contexts of operations as well. Without an executable context, such insights are not supported by using an executable architecture alone.

Sage and Rouse aligned the six key interrogatives to information and knowledge categories, distinguishing between those that relate to information and those that relate to knowledge: *who*, *what*, *where*, *when* refer to information; *how* and *why* deal with knowledge [14, p. 264]. Executable architectures should address both the information and knowledge categories. Adding system context allows us to address why a system acts as it does (and in so doing following the operational requirements of a given scenario) and how it performs its actions (in the collaboration with the other influencing systems (by meeting mission need).

All information needed to provide for the context is normally captured in systems engineering documents. The systems architecture is based on operational requirements (OR) that are derived from mission requirements (MR). These OR are refined into Systems Requirements (SR), Functional Requirements (FR), and Component Requirements (CR), which build the foundation for the systems architecture. MR and OR can be used to identify scenarios and metrics to measure the success of a mission. Figure 3 shows how the executable architecture is derived from SR, FR, and CR to be embedded into an executable context based on MR and OR.

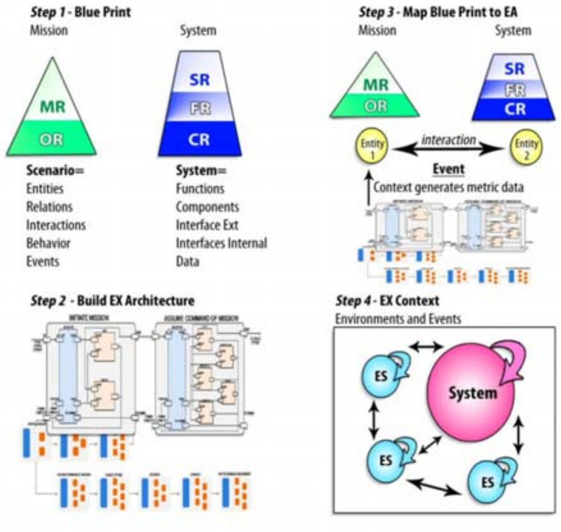


Figure 3. Executable Architecture in the Executable Context

Using the DEVS Unified Process (DUNIP) developed in [15], the system architecture is represented as an executable architecture in JAVA code and can react to inputs as defined in the system architecture and can produce the outputs using the appropriate causal and temporal constraints as defined for the systems.

Using validated simulation systems representing the context and the external systems within critical missions identified in the MR and OR, the validation of the architecture can now be conducted in the context of a valid scenario, using metrics identified by the real user for the critical missions. Garcia applied the NATO Code of Best Practice for C2 Assessment [16] and the Military Missions to Means Framework (MMF) [17].

This approach allows us to identify counter-intuitive effects, such as worst overall results.

5.0 CONCLUSION

The research currently conducted on executable architecture at Old Dominion University will contribute to better processes for validation of system development. Adding the power of M&S solutions to the rigor of systems engineering allows much better decisions on all level, from the stakeholder and future user of the system down to the implementing engineer. Executable architectures following the theory and being embedded into an executable context will allow all partners to display and evaluate operationally relevant data in agile contexts by executing models using operational data exploiting the full potential of M&S and producing numerical insight into the behavior of complex systems.

6.0 REFERENCES

[1] Adshead, S., Kreitmair, T., and Tolk, A. 2001. Definition of ALTBMD Architectures by Applying the C4ISR Architecture Framework. *Proceedings Fall Simulation Interoperability Workshop*, Vol. II, pp. 679 – 689, Orlando, FL, September

[2] Zinn, A. W. 2004. The Use of Integrated Architectures to Support Agent Based Si-

mulation An Initial Investigation. *Master's Thesis, Air Force Institute of Technology Air University*

[3] Wagenhals, L. W, Haider, S., and Levis, A. H. 2002. Synthesizing Executable Models of Object Oriented Architectures. *Proceedings of Workshop on Formal Methods Applied to Defence Systems,* Adelaide, Australia, pp. 85-93

[4] Zeigler, B. P., and Mittal, S. 2005. Enhancing DoDAF with a DEVS-Based System Lifecycle Development Process. *IEEE International Conference on Systems, Man and Cybernetics*, Hawaii, October

[5] Mittal, S. 2006, Extending DoDAF to Allow Integrated DEVS-Based Modeling and Simulation. *JDMS* 3(2):95-123

[6] Ring, S. J., Nicholson, D., and Pallab S. 2007. Activity-Based Methodology for Development and Analysis of Integrated DoD Architectures. *Information Science Reference: Handbook of Enterprise Systems Architecture in Practice*, Chapter 5, pp. 85-113, Systems Modeling Language OMG, 2008

[7] Mittal, S., Mitra, A., Gupta, A, and Zeigler, B.P. 2006. Strengthening OV-6a Semantics with Rule-Based Meta-models in DEVS/DoDAF based Life-cycle Architectures Development. *IEEE-Information Reuse and Integration*, Special Section on DoDAF, Hawaii

[8] Risco-Martin, J.L., de la Cruz, J., Mittal, S., and Zeigler, B.P. 2009. Eudevs: Executable UML with DEVS Theory of Modeling and Simulation, *Simulation* 85(7):419-450

[9] Shuman, E. A. 2010. Understanding Executable Architectures Through An Examination of Language Model Elements. *SCS Proceedings of the Summer Computer Simulation Conference*, Ottawa, Canada, July

[10] Fishwick, P. 1995. *Simulation Model Design and Execution.* Prentice-Hall, Inc.

[11] Garcia, J. J., and Tolk, A. 2010. Adding Executable Context to Executable Architectures: Shifting Towards a Knowledge-Based Validation Paradigm for System-of-Systems Architectures. *SCS Proceedings of the Summer Computer Simulation Conference*, Ottawa, Canada, July

[12] Garcia, J. J. 2010. Methodology Supporting Architecture Validations (MAVS). *SCS Proceedings of the Spring Simulation Multi-Conference*, Symposium Emerging Applications of M&S in Industry and Academia, Orlando, FL, April

[13] Buede, D. 2000. *The Engineering Design of Systems: Models and Methods*, John Wiley & Sons, Inc., New York

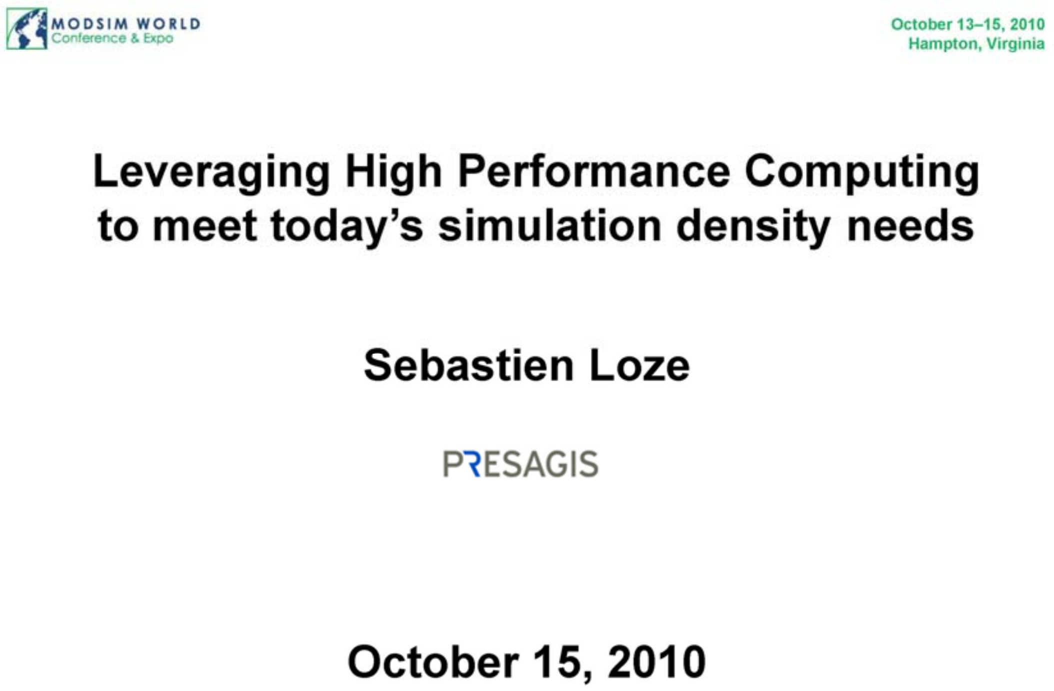
[14] Sage, A. P., and Rouse, W. B. (Eds.). 1999. *Handbook of Systems Engineering and Management*, John Wiley and Sons, Inc., New York

[15] Mittal S. 2007. DEVS unified process for integrated development and testing of service oriented architectures. *PhD Thesis. United States -- Arizona: The University of Arizona*

[16] *NATO Code of Best Practice for C2 Assessment*, 2002, CCRP Press, Washington DC

[17] Deitz, P.H., Sheehan, J.H., Harris, B.A., Wong, A.B.H., Bray, B.E., and Purdy, E.M. 2003. The Military Missions and Means Framework (MMF). *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC),* Orlando, FL, December

2.2 Leveraging High Performance Computingto meet today’s simulation density needs



MODSIM WORLD
Conference & Expo

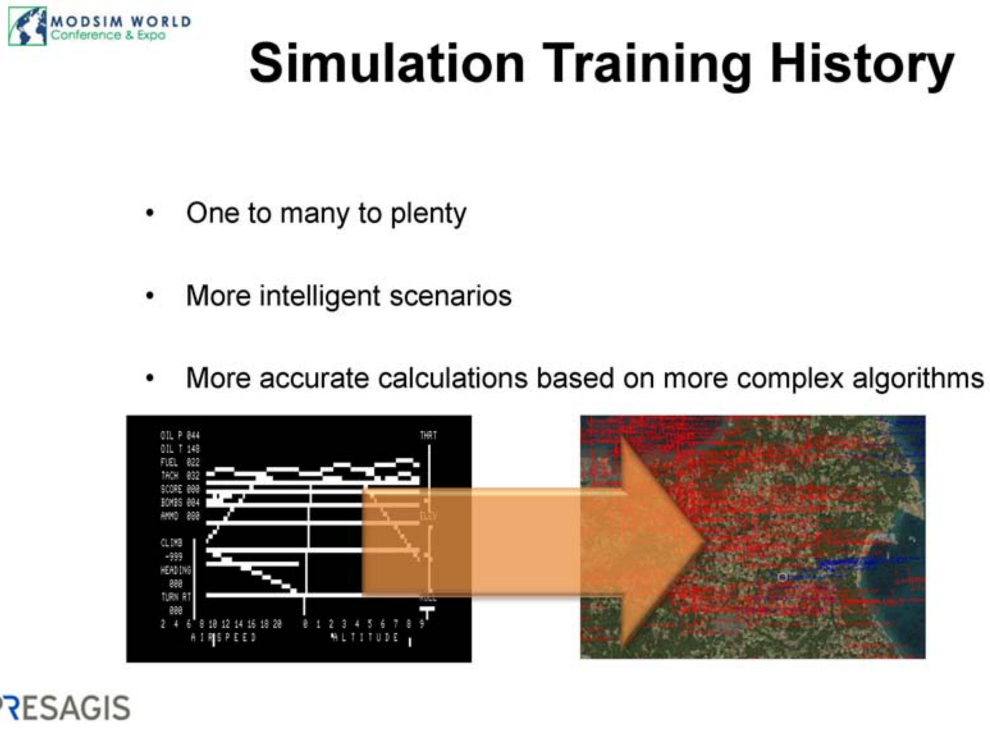
October 13–15, 2010
Hampton, Virginia

Leveraging High Performance Computing to meet today’s simulation density needs

Sebastien Loze

PRESAGIS

October 15, 2010

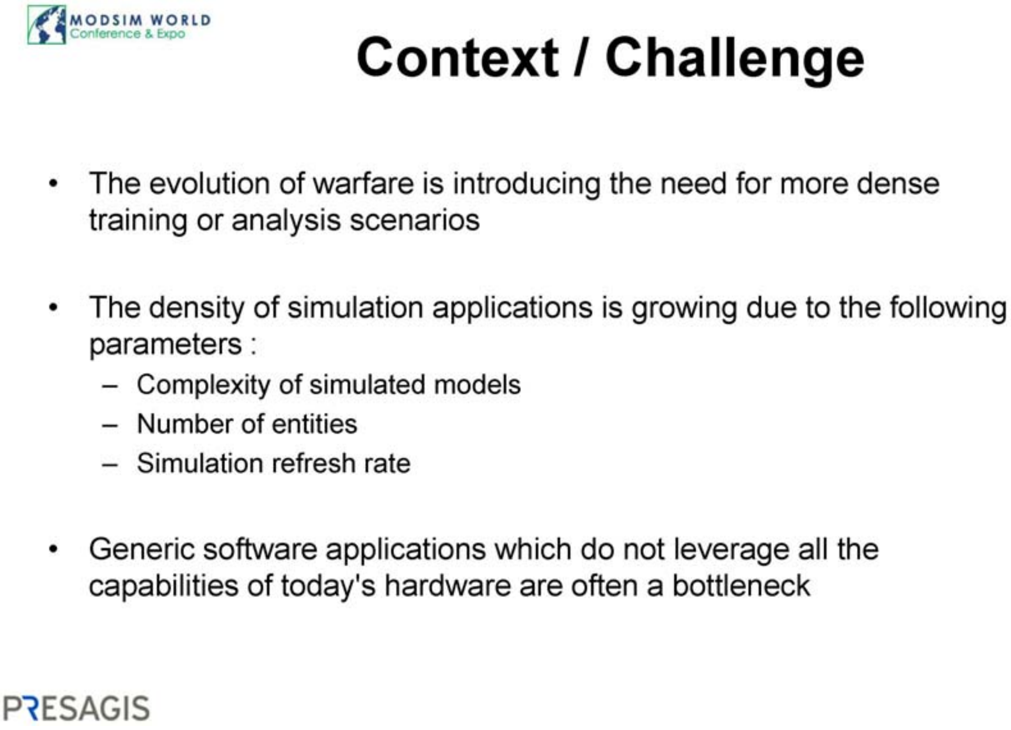


MODSIM WORLD
Conference & Expo

Simulation Training History

- One to many to plenty
- More intelligent scenarios
- More accurate calculations based on more complex algorithms

PRESAGIS

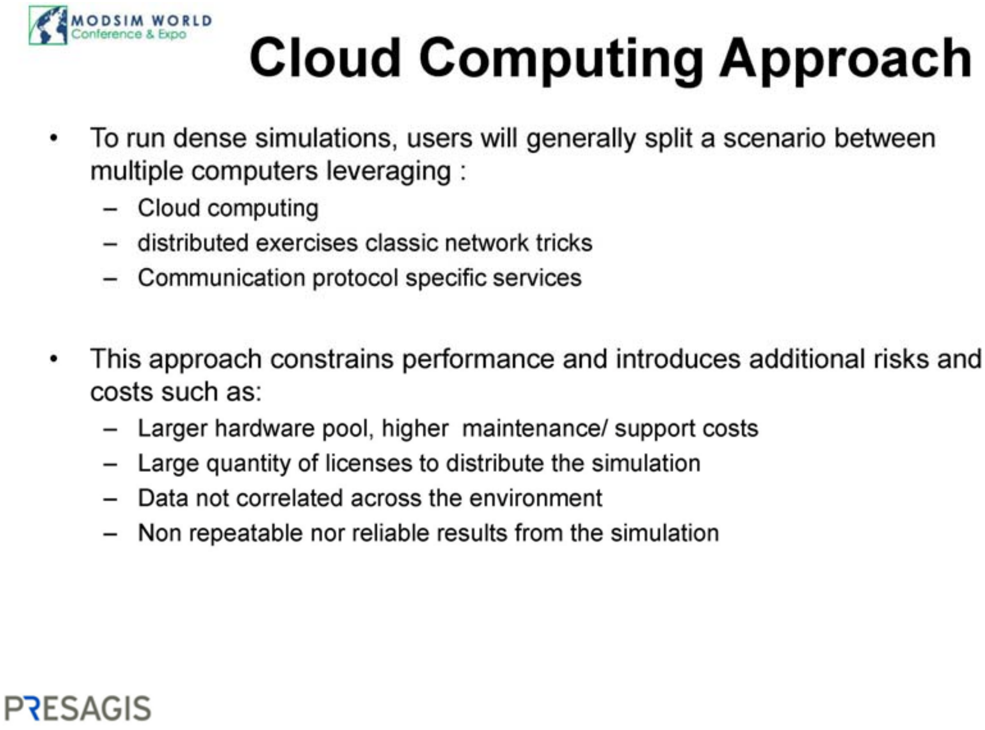


MODSIM WORLD Conference & Expo

Context / Challenge

- The evolution of warfare is introducing the need for more dense training or analysis scenarios
- The density of simulation applications is growing due to the following parameters :
  - Complexity of simulated models
  - Number of entities
  - Simulation refresh rate
- Generic software applications which do not leverage all the capabilities of today's hardware are often a bottleneck

PRESAGIS

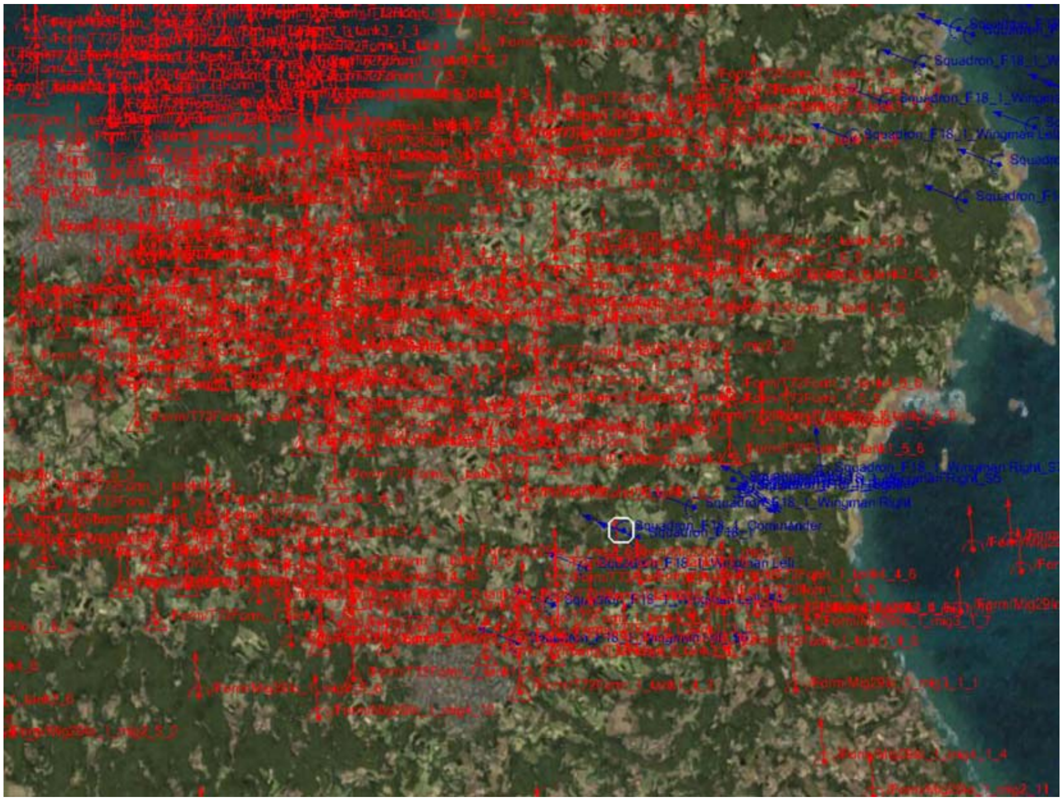


MODSIM WORLD Conference & Expo

Cloud Computing Approach

- To run dense simulations, users will generally split a scenario between multiple computers leveraging :
  - Cloud computing
  - distributed exercises classic network tricks
  - Communication protocol specific services
- This approach constrains performance and introduces additional risks and costs such as:
  - Larger hardware pool, higher maintenance/ support costs
  - Large quantity of licenses to distribute the simulation
  - Data not correlated across the environment
  - Non repeatable nor reliable results from the simulation

PRESAGIS



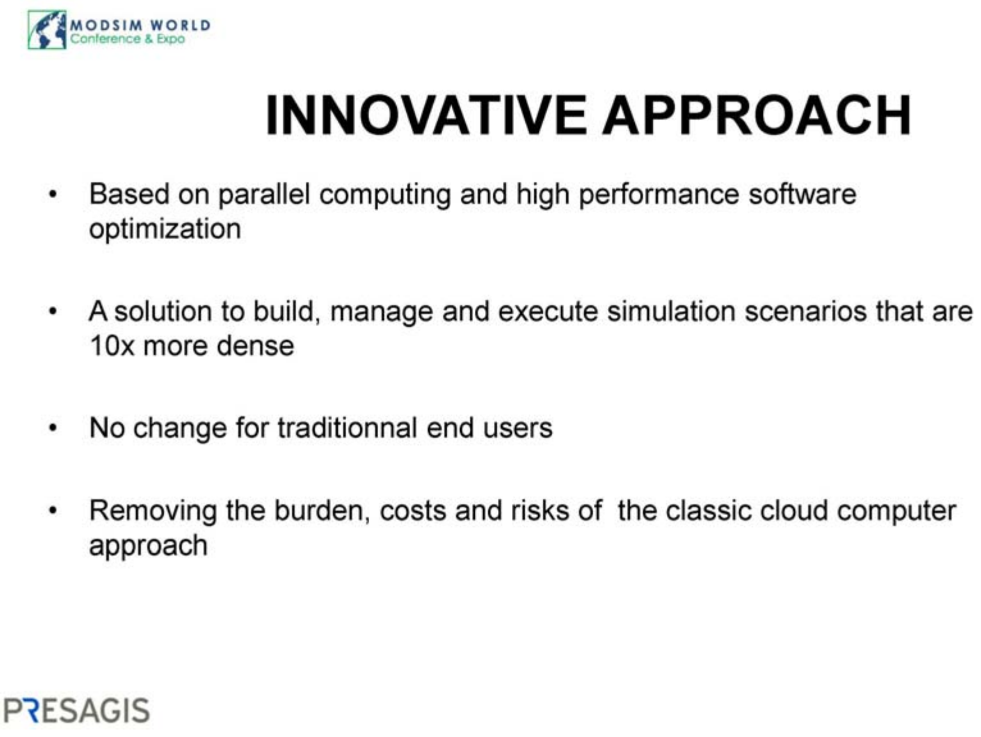
MODSIM WORLD Conference & Expo

COTS Simulation tool

- Powerful, open and flexible simulation software
- Already a proven and adopted solution to simulate thousands of complex entities at real-time rates with no frame overruns, leveraging distributed exercise techniques

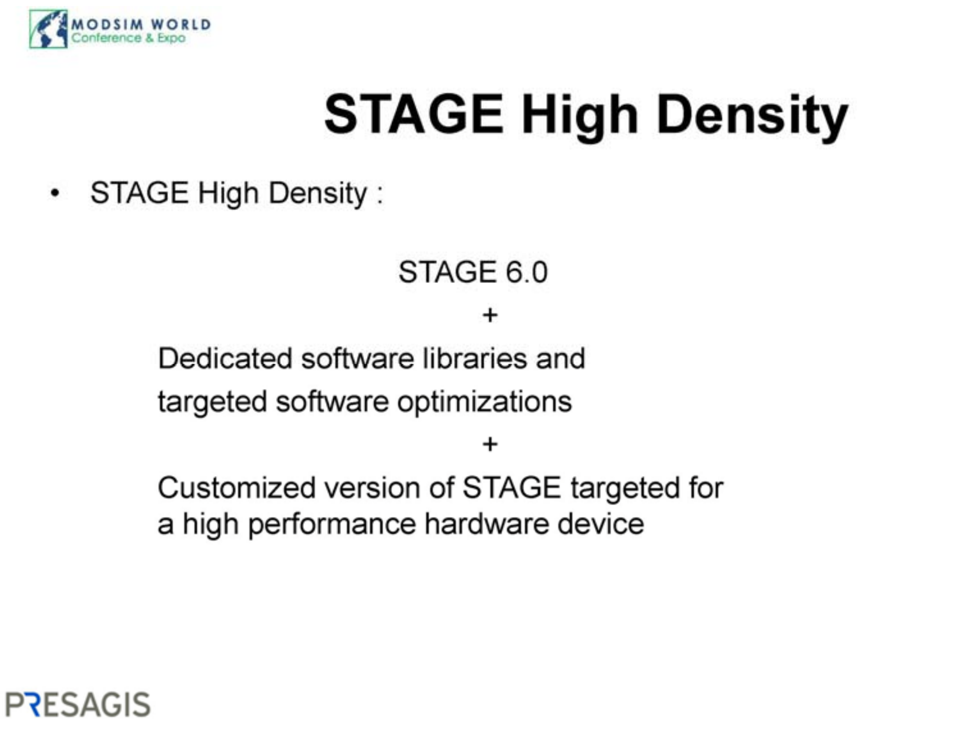
"STAGE can handle on the order of 10,000 entities.
This is the going requirement to support VF-size events."
Major d'Artagnan R. de Anda
Chief of the Distributed Warfare Center

PRESAGIS



INNOVATIVE APPROACH

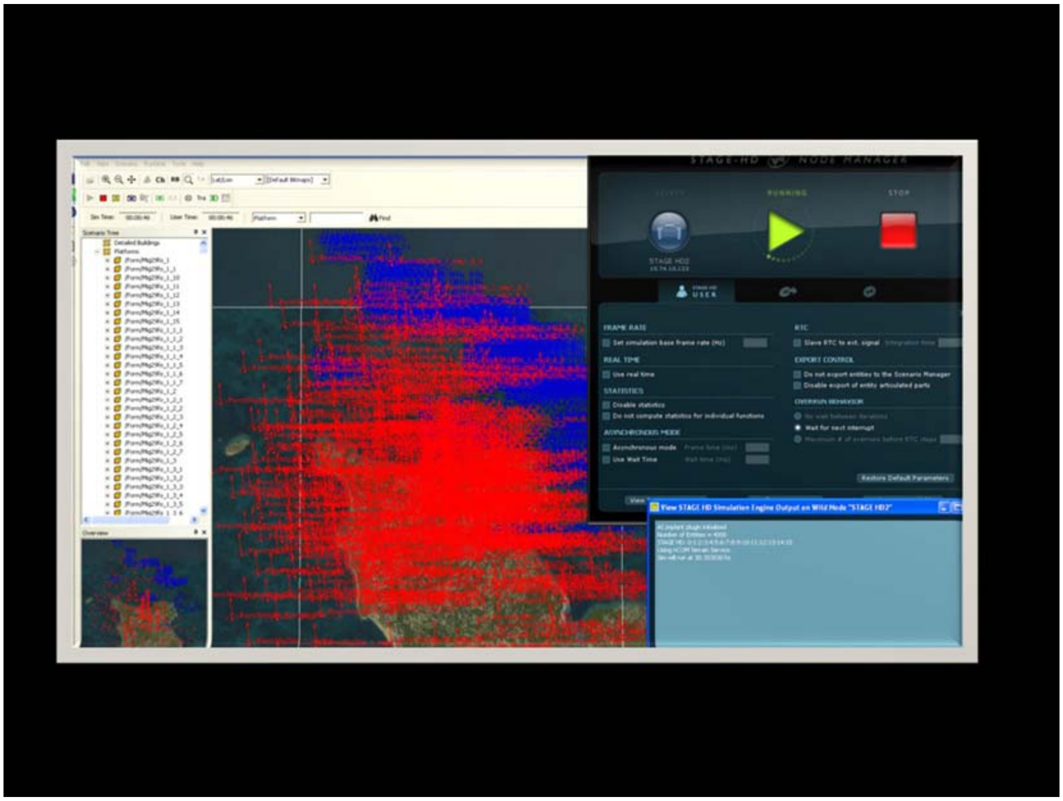
- Based on parallel computing and high performance software optimization
- A solution to build, manage and execute simulation scenarios that are 10x more dense
- No change for traditionnal end users
- Removing the burden, costs and risks of the classic cloud computer approach



STAGE High Density

- STAGE High Density :

STAGE 6.0
+
Dedicated software libraries and
targeted software optimizations
+
Customized version of STAGE targeted for
a high performance hardware device



FRAME RATE

□ Set simulation base frame rate (Hz)

REAL TIME

□ Use real time

STATISTICS

□ Disable statistics

□ Do not compute statistics for individual functions

ASYNCHRONOUS MODE

□ Asynchronous mode Frame time (ms)

□ Use Wait Time Wait time (ms)

RTC

□ Slave RTC to ext. signal Integration time

EXPORT CONTROL

□ Do not export entities to the Scenario Manager

□ Disable export of entity articulated parts

OVERRUN BEHAVIOR

○ No wait between iterations

● Wait for next interrupt

○ Maximum # of overruns before RTC steps

View STAGE HD Simulation Engine Output on Wild Node "STAGE HD2"

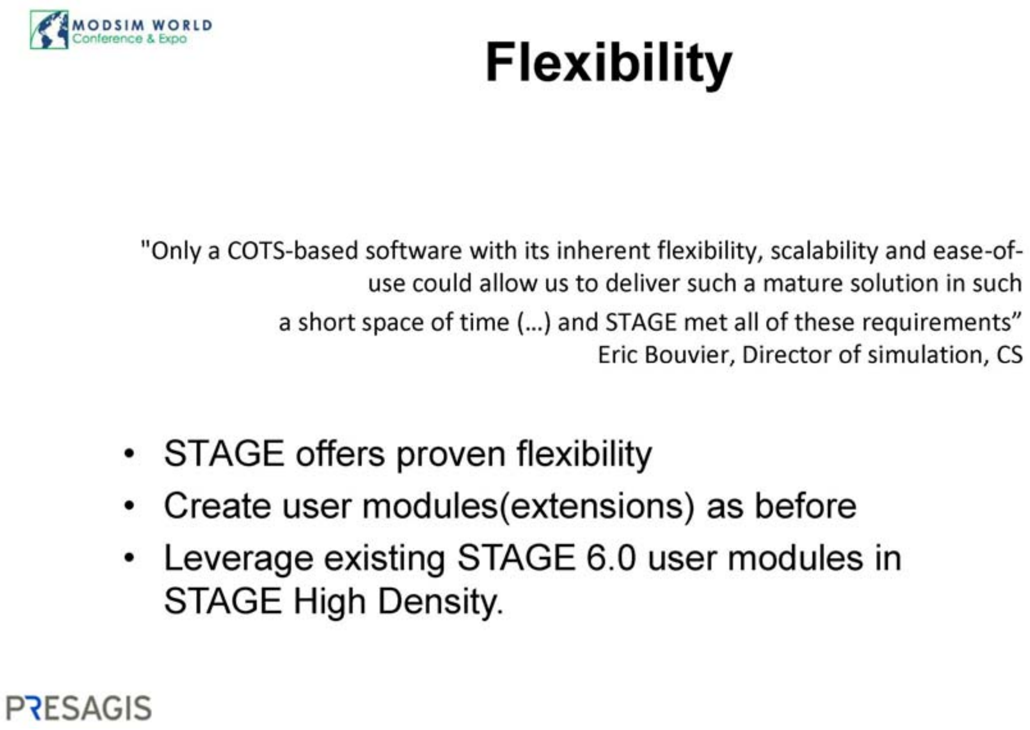
```
AI implant plugin initialized
Number of Entities = 4000
STAGE HD: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Using HCOM Terrain Service
Sim will run at 30.303030 Hz
```

MODSIM WORLD Conference & Expo

Performance goals

- Create and run scenarios that are 10x more dense
  - higher fidelity simulation models
  - larger-scale simulation exercises
  - Limited frame overrun with same simulation rate
  - More model updates in asynchronous mode

PRESAGIS



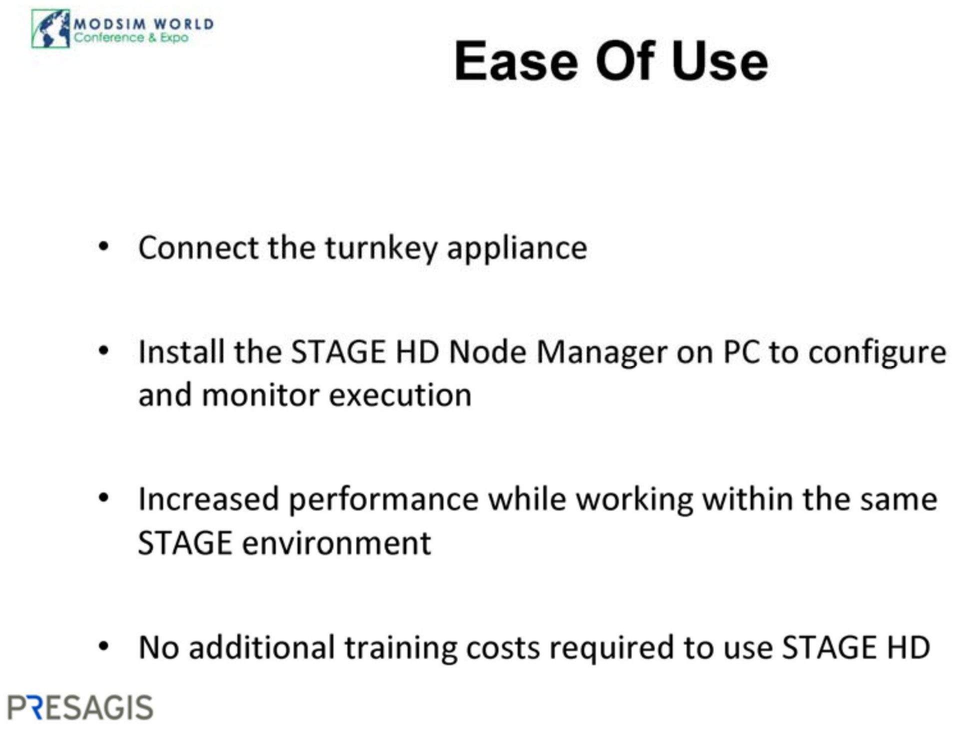
MODSIM WORLD Conference & Expo

Flexibility

"Only a COTS-based software with its inherent flexibility, scalability and ease-of-use could allow us to deliver such a mature solution in such a short space of time (…) and STAGE met all of these requirements"
Eric Bouvier, Director of simulation, CS

- STAGE offers proven flexibility
- Create user modules(extensions) as before
- Leverage existing STAGE 6.0 user modules in STAGE High Density.

PRESAGIS

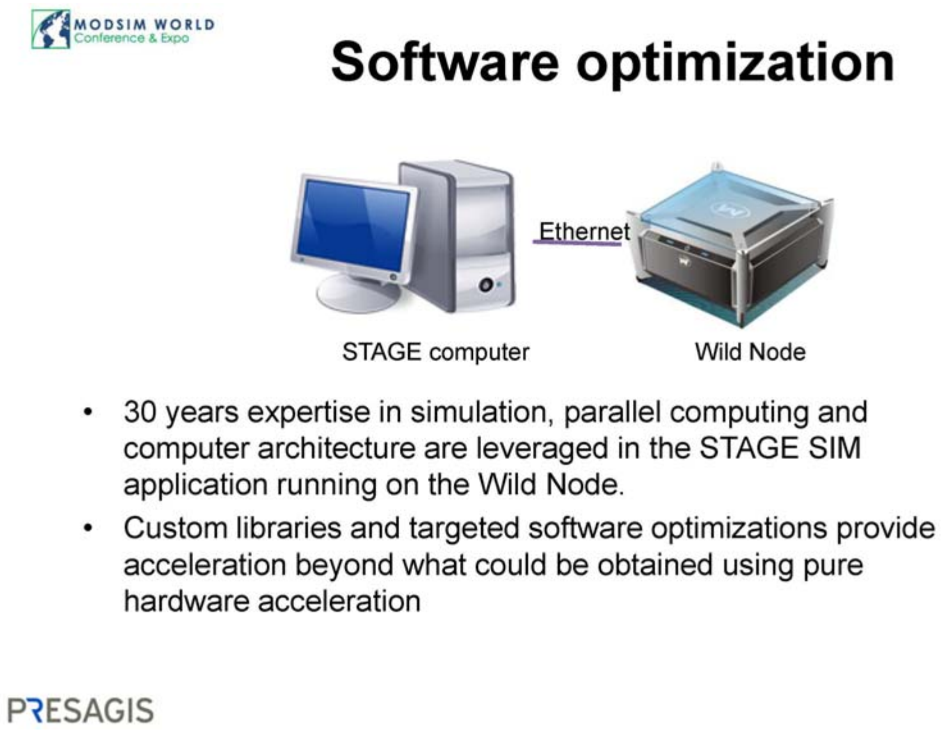


MODSIM WORLD Conference & Expo

Ease Of Use

- Connect the turnkey appliance
- Install the STAGE HD Node Manager on PC to configure and monitor execution
- Increased performance while working within the same STAGE environment
- No additional training costs required to use STAGE HD

PRESAGIS



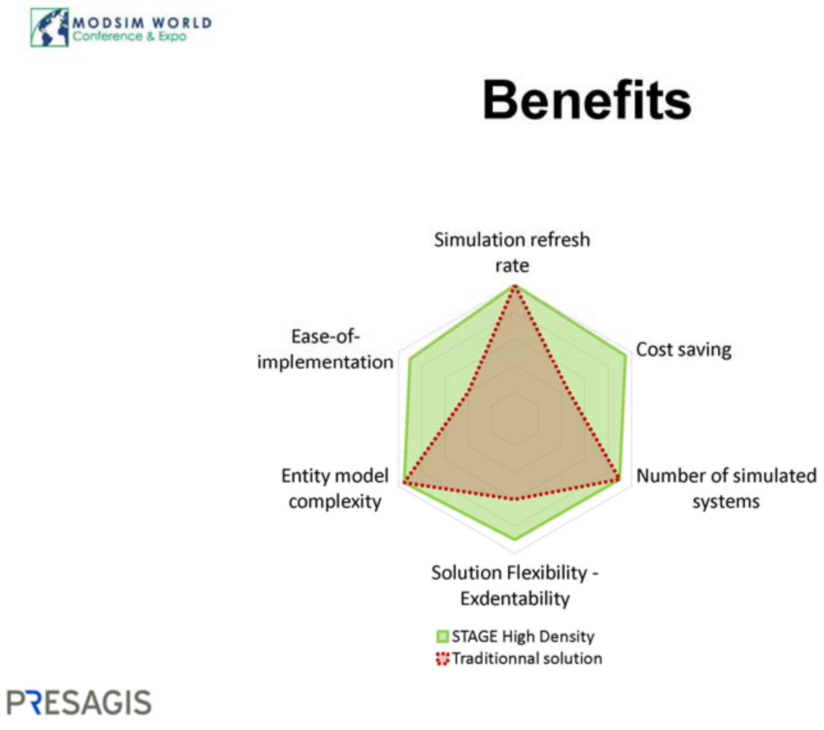
MODSIM WORLD Conference & Expo

Software optimization

Ethernet
STAGE computer
Wild Node

- 30 years expertise in simulation, parallel computing and computer architecture are leveraged in the STAGE SIM application running on the Wild Node.
- Custom libraries and targeted software optimizations provide acceleration beyond what could be obtained using pure hardware acceleration

PRESAGIS



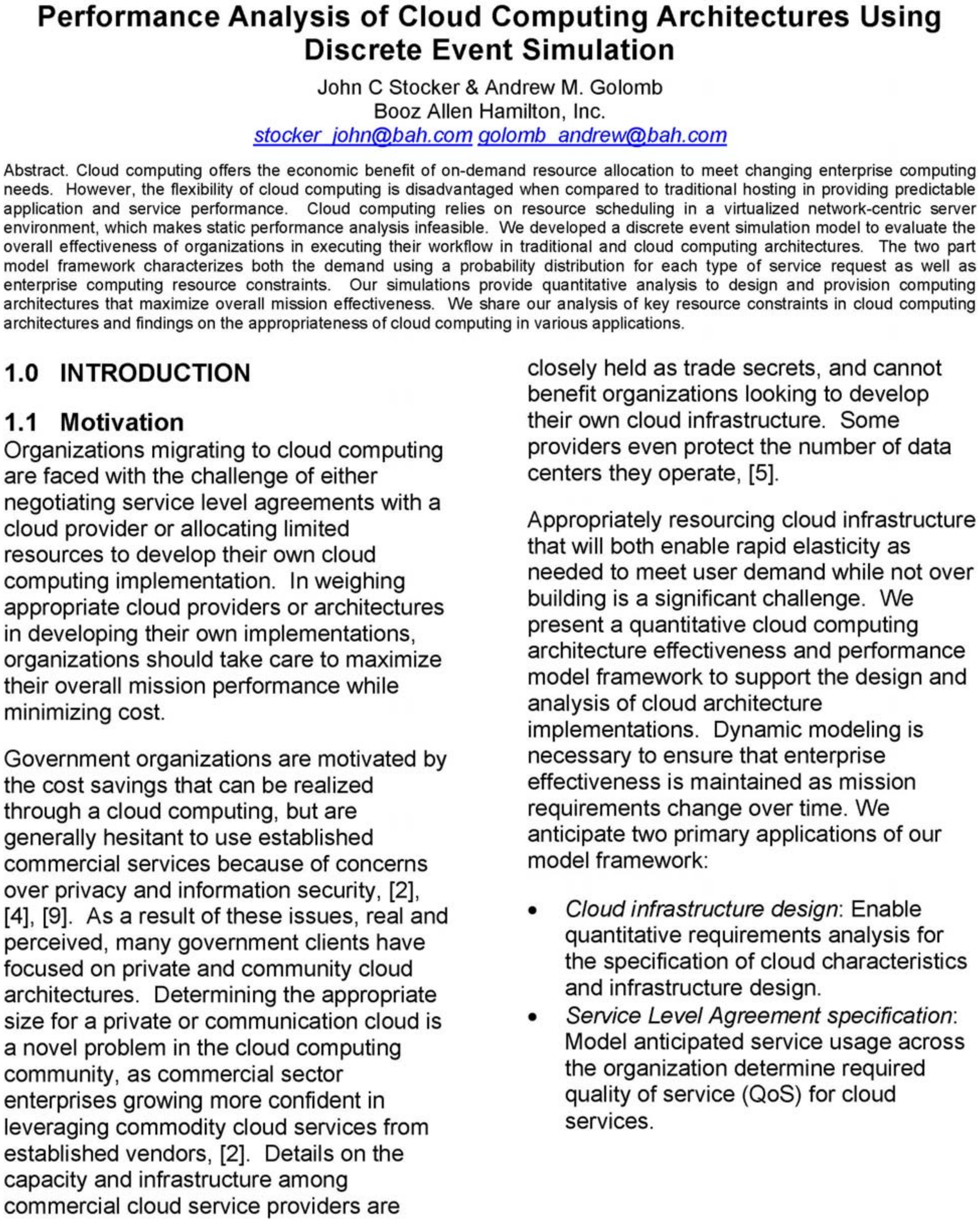
MODSIM WORLD Conference & Expo

Benefits

PRESAGIS

Power
Wild Node
Reset

2.3 Performance Analysis of Cloud Computing Architectures Using Discrete Event Simulation



Performance Analysis of Cloud Computing Architectures Using Discrete Event Simulation

John C Stocker & Andrew M. Golomb
Booz Allen Hamilton, Inc.
stocker_john@bah.com golomb_andrew@bah.com

Abstract. Cloud computing offers the economic benefit of on-demand resource allocation to meet changing enterprise computing needs. However, the flexibility of cloud computing is disadvantaged when compared to traditional hosting in providing predictable application and service performance. Cloud computing relies on resource scheduling in a virtualized network-centric server environment, which makes static performance analysis infeasible. We developed a discrete event simulation model to evaluate the overall effectiveness of organizations in executing their workflow in traditional and cloud computing architectures. The two part model framework characterizes both the demand using a probability distribution for each type of service request as well as enterprise computing resource constraints. Our simulations provide quantitative analysis to design and provision computing architectures that maximize overall mission effectiveness. We share our analysis of key resource constraints in cloud computing architectures and findings on the appropriateness of cloud computing in various applications.

1.0 INTRODUCTION

1.1 Motivation

Organizations migrating to cloud computing are faced with the challenge of either negotiating service level agreements with a cloud provider or allocating limited resources to develop their own cloud computing implementation. In weighing appropriate cloud providers or architectures in developing their own implementations, organizations should take care to maximize their overall mission performance while minimizing cost.

Government organizations are motivated by the cost savings that can be realized through a cloud computing, but are generally hesitant to use established commercial services because of concerns over privacy and information security, [2], [4], [9]. As a result of these issues, real and perceived, many government clients have focused on private and community cloud architectures. Determining the appropriate size for a private or communication cloud is a novel problem in the cloud computing community, as commercial sector enterprises growing more confident in leveraging commodity cloud services from established vendors, [2]. Details on the capacity and infrastructure among commercial cloud service providers are closely held as trade secrets, and cannot benefit organizations looking to develop their own cloud infrastructure. Some providers even protect the number of data centers they operate, [5].

Appropriately resourcing cloud infrastructure that will both enable rapid elasticity as needed to meet user demand while not over building is a significant challenge. We present a quantitative cloud computing architecture effectiveness and performance model framework to support the design and analysis of cloud architecture implementations. Dynamic modeling is necessary to ensure that enterprise effectiveness is maintained as mission requirements change over time. We anticipate two primary applications of our model framework:

- *Cloud infrastructure design*: Enable quantitative requirements analysis for the specification of cloud characteristics and infrastructure design.
- *Service Level Agreement specification*: Model anticipated service usage across the organization determine required quality of service (QoS) for cloud services.

1.2 Analyzing Cloud Effectiveness and Performance

We developed a cloud computing effectiveness and performance model framework to analyze the effectiveness of various cloud computing architectures in meeting enterprise computing requirements by applying operations research techniques. The analytic model coupled with a companion cost analysis model [1] will enable the development of comprehensive migration strategies that right-size investments in developing private, community, and public cloud computing solutions. While previous research has analyzed the performance of cloud computing [10], that work focused on cloud service requirements, and not detailed analysis of various cloud architectures. To the best of our knowledge, we are the first to develop a generalized approach that can provide analysis of detailed cloud infrastructure characteristics.

2.0 EFFECTIVENESS AND PERFORMANCE MODEL

The effectiveness and performance model combines a dynamic workforce business process model with a cloud computing architecture resource model. Figure 1 demonstrates a top level view of model components. The model framework can be configured to simulate specific processes and a variety of cloud environments. Discrete event simulation is used to execute business processes in the context of cloud computing architectures. The model is based on a discrete event simulation model built using ExtendSim® [3] developed by Imagine That Inc.

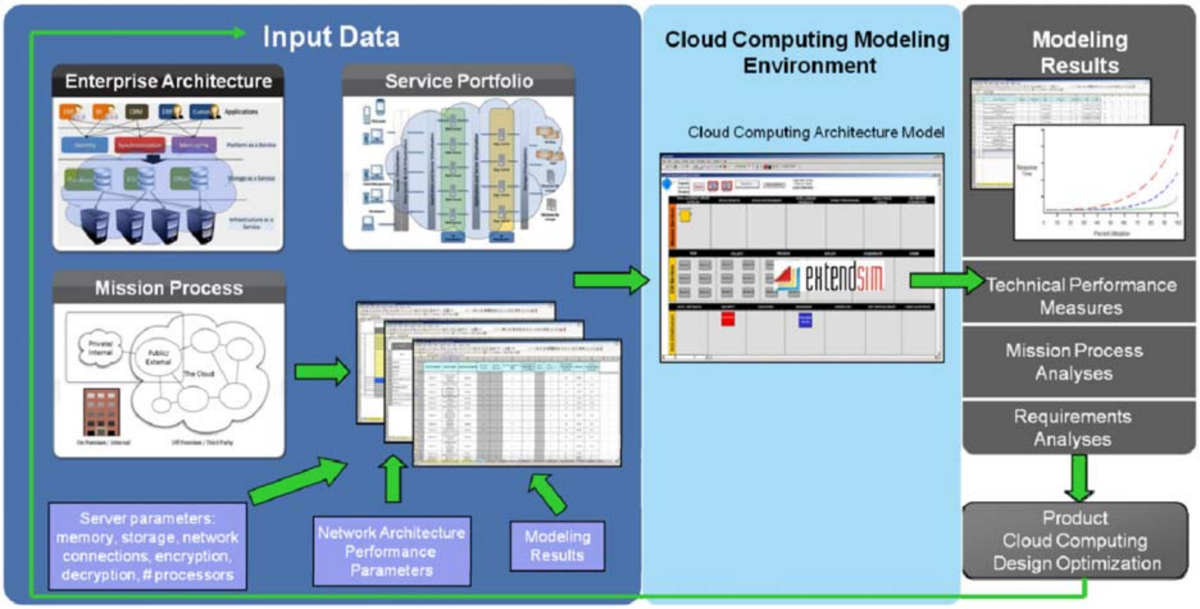

Input Data
Enterprise Architecture
Service Portfolio
Mission Process
Server parameters: memory, storage, network connections, encryption, decryption, # processors
Network Architecture Performance Parameters
Modeling Results
Cloud Computing Modeling Environment
Cloud Computing Architecture Model
extendsim
Modeling Results
Technical Performance Measures
Mission Process Analyses
Requirements Analyses
Product Cloud Computing Design Optimization

Figure 1. Cloud Computing Modeling Methodology

2.1 Business Process Model

The business process model generates compute service requests that are dispatched to the cloud for execution. The process model considers the number of users, their roles, and the distribution of service requests based on their roles. Services are configured to reflect the organization's mission and can include generic and specific activities such as: email, instant messaging, video chats, web browsing, document review and composition, budget and financial analysis, and scientific computing.

The business process model is currently comprised of two top level models:

- Office
- Service

Numbers of staff are affiliated within offices to model the overall organization effectiveness in community cloud architectures. Network characteristics between users and cloud computing resources are specified at the office level to accurately model differences in network connectivity between users of community clouds. Table 1 describes the characteristics of the office model:

Office site name
Number of staff members
Communication link characteristics - Maximum uplink bandwidth (Mbps) - Maximum downlink bandwidth (Mbps)

Table 1. Office model

The business process model has sophisticated capability to generate service demand based on the rhythm in the workforce and enterprise needs. The process model can also represent surges in service demand that may be triggered by mission events. This capability is crucial in analyzing the benefits of cloud elasticity in meeting mission requirements. Table 2 defines the characteristics of the software or service modeled.

Application or service name
Request rate
Data transferred (KB)
Computational requirements (MIPS)
QoS requirements - Allowable response latency

Table 2. Service model

The service model characterizes request rate in terms of on demand, hourly, daily, weekly, monthly, and quarterly. On demand requests are specified using a statistical distribution of demand rate or interarrival time. Data transferred and computational requirements are also statistical distributions that define average and peak values.

Using statistical distributions permits significant flexibility in modeling various services and processes. Distributions can be varied (e.g., Gaussian, Poisson, triangular) to most accurately represent the specific enterprise and mission being modeled. The business process model framework is flexible and can represent diverse work streams throughout the enterprise.

2.2 Cloud Computing Architecture Model

The cloud computing architecture model captures overall cloud configuration and resources. The model has a flexible design that facilitates simulation of various cloud computing architecture deployment models identified by the National Institute of Standards and Technology [6], also shown in Figure 2 [7].

- Community
- Hybrid
- Private
- Public

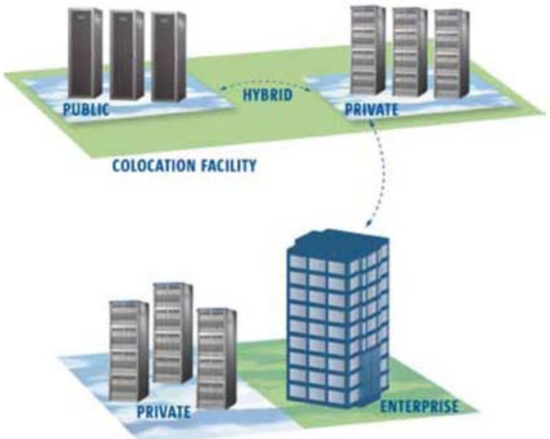


Figure 2. Cloud Computing Architectures

The cloud computing architecture model represents the underlying cloud infrastructure, including network and servers. The model assumes uniform performance with respect to the three cloud service models: Software as a Service (SaaS), Platform as a Service (Paas), and Infrastructure as a Service (IaaS). The distinctions between these service models are largely based on revenue models and not technical distinctions that affect overall performance of a cloud-based service solution.

NIST further describes the characteristics of cloud computing environments, including:

- On-demand self-service
- Broad network access
- Resource pooling
- Rapid elasticity
- Measured service

The scope of model focuses on providing an analysis of necessary cloud computing node resource characteristics, such as network access, processing capacity, memory, and storage. The model assumes that the nodes are homogenous. The internal network resources between cloud nodes are also modeled and can represent various server interconnectivity solutions. Table 3 specifies the characteristics captured in the cloud architecture model.

Number of Nodes
Node processing capacity
Node memory
Node storage capacity
Node storage access speed
Internode network characteristics

Table 3. Cloud model

A limitation of the cloud architecture model in its current state is its inability to represent hybrid cloud solutions that incorporate multiple instances of cloud environments into a single implementation. This limitation restricts current analysis to consider only private, community, and public infrastructure. Future work is planned to incorporate the ability to represent hybrid implementations that consider combinations of private, community, and public cloud environments.

It is important to note that the cloud architecture model framework is also flexible enough to represent traditional fixed hosting infrastructure to permit comparisons of service performance between cloud and traditional hosting environments.

3.0 EFFECTIVENESS AND PERFORMANCE METRICS

3.1 Measures of Effectiveness

Measures of effectiveness focus on the ability of the cloud infrastructure to meet service demand in a timely fashion and enable the organization to perform its mission.

3.1.1 On-Time Response Rate

On-time response rate is the cumulative distribution of service requests that are completed and responded to within the required latency. Acceptable latency is service and mission specific and typically varies in magnitude from fractions of seconds to minutes.

3.1.2 Service Latency

Service latency is the delay experienced by users in processing service demand. This measure is an indicator of services along the critical path in mission execution and an indicator of potential idleness in the workforce.

3.2 Measures of Performance

Measures of performance characterize cloud resource utilization in meeting the service demand. The measures of performance gauge cloud elasticity in the context of an enterprise's mission requirements.

3.2.1 Cloud Processing Utilization

Cloud processing utilization, as measured by the usage of overall cloud processing capacity over time and across all services in the mission workflow, is an indication of overall processing demand.

3.2.2 Cloud Storage Utilization

Cloud storage utilization, as measured by the usage of cloud storage capacity over time and across the workflow, is an indication of overall storage demand.

3.2.3 Cloud Communications Utilization

Cloud communication utilization is the percentage of communications resources over time. Individual measures are recorded for each office communications link as well as the cloud communications infrastructure.

4.0 ANALYSIS

Quantitative cloud architecture analysis using the effectiveness and performance model will help answer fundamental design questions and support trade studies that refine specific cloud technologies and solution sets. Analyzing the impact of cloud architectures characteristics on the enterprise is crucial to designing specifying acceptable cloud architecture implementations.

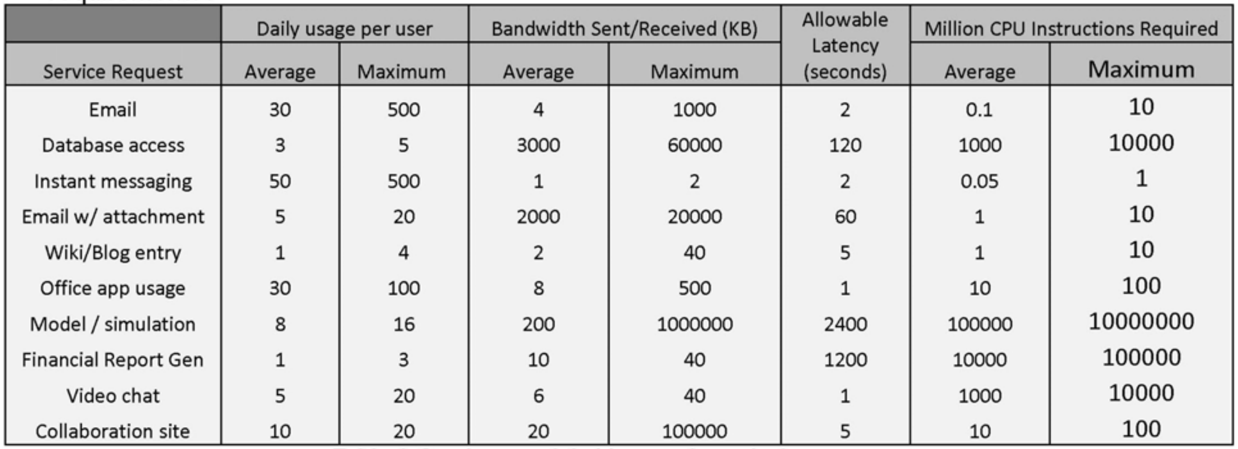
The cloud computing effectiveness and performance model will help answer design questions such as:

- Is a cloud architecture appropriate for the mission?
- Is a private, community, or public cloud best suited to meet performance requirements?
- Does the cloud have sufficient elasticity for all enterprises, offices, users, and missions?
- What is the right size cloud to meet mission requirements and minimize cost?
- Do communications resources need to be upgraded when services are migrated to a cloud architecture?

These questions are best addressed by considering the overall on-time response rate across all service requests among all users relative to the performance metrics. Ideally, the on-time response rate is 100%, however, given the scale of requests across the enterprise and resulting peaks of activity it's impractical to build a cloud that can scale rapidly enough to meet 100% on-time response rate at all times.

4.1 Example sensitivity analysis

We conducted sensitivity analysis of on-time response rate based on some of the key cloud characteristics for a sample private cloud infrastructure. The sensitivity analysis of on-time response rate was performed by varying cloud computing architecture resource parameters. We looked for deflection points where there were diminishing returns in the overall on-time response rate relative to additional cloud resources: either more additional nodes or increased bandwidth.



	Daily usage per user		Bandwidth Sent/Received (KB)		Allowable Latency (seconds)	Million CPU Instructions Required	
Service Request	Average	Maximum	Average	Maximum		Average	Maximum
Email	30	500	4	1000	2	0.1	10
Database access	3	5	3000	60000	120	1000	10000
Instant messaging	50	500	1	2	2	0.05	1
Email w/ attachment	5	20	2000	20000	60	1	10
Wiki/Blog entry	1	4	2	40	5	1	10
Office app usage	30	100	8	500	1	10	100
Model / simulation	8	16	200	1000000	2400	100000	10000000
Financial Report Gen	1	3	10	40	1200	10000	100000
Video chat	5	20	6	40	1	1000	10000
Collaboration site	10	20	20	100000	5	10	100

Table 4. Services modeled in sample analysis

Our example sensitivity analysis considers a single enterprise office with 5,000 staff members using a private cloud infrastructure. In this sample sensitivity analysis we assumed that all users had uniform workflow and used services in equal distributions. It is reasonable to approximate usage at the service level across the enterprise instead of the user role level. Table 4 provides a summary ofthe service characteristics modeled in this example analysis.

Table 5 describes the cloud architecture characteristics used in this example analysis. Given the private cloud studied in this example, the bandwidth is the data rate between the office and the cloud implementation.

Bandwidth (Mbps)	5,000
Number of nodes	500
CPU Processor speed (GHz)	2
Instructions per cycle	16
Storage capacity per node (TB)	0.1
RAM (GB)	1

Table 5. Cloud architecture modeling in sample analysis

In our sensitivity analysis we independently varied the available bandwidth between the office and cloud and the number of nodes to assess the impact of each on the on-time response rate. Other variables were held at values that would not impact effectiveness. In the sensitivity analysis, the cloud architecture was fixed at 500 nodes.

4.2 Example sensitivity analysis

We observed a high sensitivity on bandwidth in our example analysis. Figure 3 shows the sensitivity of the on-time response rate on bandwidth.

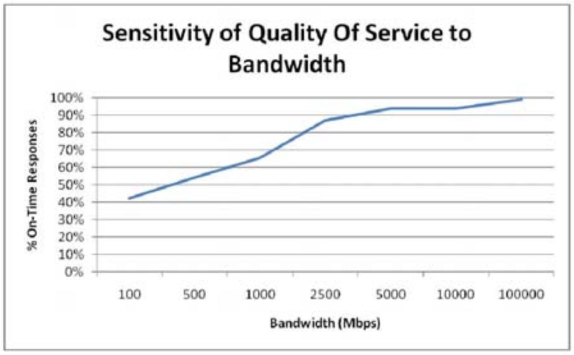


Figure 3. Bandwidth sensitivity sample analysis

The on-time response rate shown varies from approximately 42% to 94% as bandwidth is increased from 100 Mbps to 5,000 Mbps. These results are not surprising. The services modeled were indeed activities that required a significant amount of data to be transferred between users and the cloud. Figure 4 provides additional insight on the response rate for categories of services.

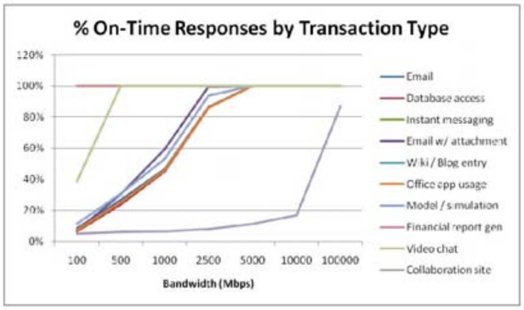


Figure 4. Bandwidth sensitivity by service

5.0 CONCLUSION AND FUTURE WORK

We have developed an effectiveness and performance model framework to assess the impact of various cloud infrastructure characteristics on overall enterprise effectiveness.

Our preliminary analysis indicates that cloud architectures may require the investment of network infrastructure to provide high quality of service to applications where a significant amount of data must be transferred.

Additional effectiveness considerations to model in the future include assessing cloud availability and analyzing various cost options in maximizing availability through large-scale implementations of commodity hardware versus more deploying limited quantities of more expensive hardware.

In the future we plan to add to the fidelity of the business process model by considering the categorization of staff members by role and modeling service request generation on a role-by-role basis. We also plan to incorporate a hierarchical business process model in which required capabilities can be decomposed into dependent software applications and services. We will leverage existing work in modeling business processes using industry standard representations such as the Federal Enterprise Architecture Framework (FEAF), [8].

The cloud computing effective and performance model provides a quantitative approach to right-size private and community cloud investments. Coupling the cloud effectiveness model with the existing Booz Allen cost model [1] will provide an overarching analytical approach to shape cloud migration strategies. This integrated analytic capability would provide detailed performance and cost analysis of an organization's mission operations in a variety of cloud environments.

6.0 REFERENCES

[1] T. Alford, G. Morton, "The Economics of Cloud Computing: Addressing the Benefits of Infrastructure in the Cloud," Booz Allen Hamilton, McLean, Virginia, 2009. [Online]. Available: http://www.boozallen.com/media/file/Economics-of-Cloud-Computing.pdf

[2] Gartner Research, "Gartner Says Worldwide Cloud Services Market to Surpass $68 Billion in 2010," *Gartner Research*, June 22, 2010. [Online] Available: http://www.gartner.com/it/page.jsp?id=1389313

[3] D. Krahl, "The Extend simulation environment," *Simulation Conference, 2001. Proceedings of the Winter* , vol.1, pp.217-225, 2001.

[4] N. Leavitt, "Is Cloud Computing Really Ready for Prime Time?," *Computer*, vol. 42, no. 1, pp. 15-20, Jan. 2009.

[5] A. Li, X. Yang, S. Kandula, M. Zhang, "CloudCmp: Shopping for a Cloud Made Easy," in *2nd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud)*, June 2010.

[6] P. Mell, T. Grance "The NIST Definition of Cloud Computing." National Institutes of Standards and Technology [6]. Version 15. 7 October 2009. [Online] Available: http://csrc.nist.gov/groups/SNS/cloud-computing/cloud-def-v15.doc

[7] Sun Microsystems. *Introduction to Cloud Computing Architecture*, 1st edition, June 2009.

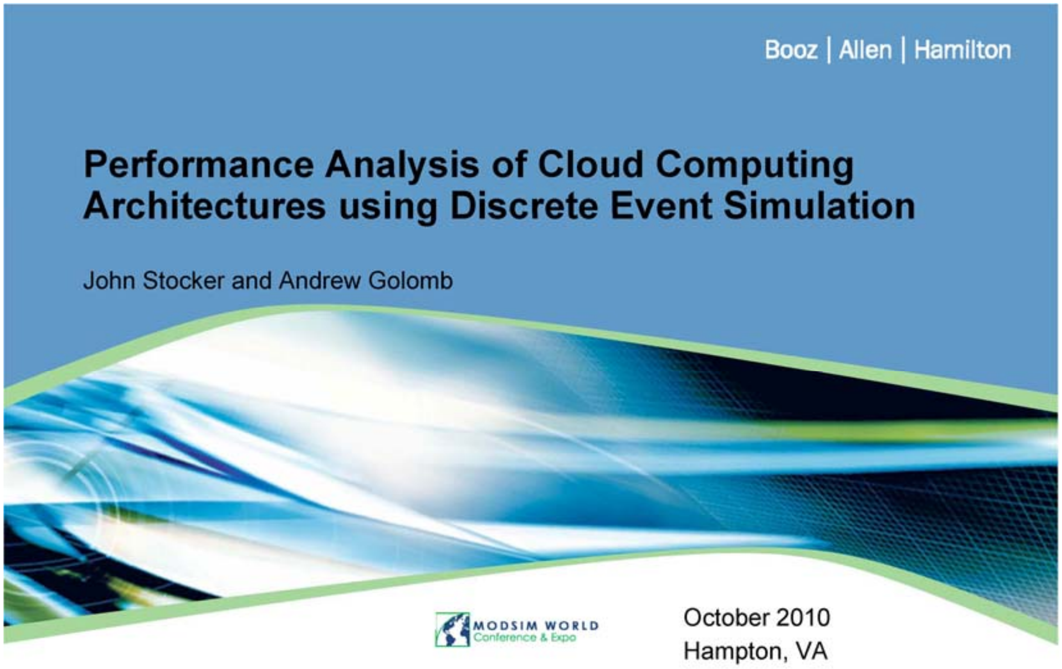
[8] United States. The Chief Information Officers Council. *Federal Enterprise Architecture Framework*, version 1.1, September 1999. [Online] Available: www.cio.gov/documents/fedarch1.pdf

[9] United States. Federal Chief Information Officer. *State of Public Sector Cloud Computing*. May 20, 2010.

[10] Kaiqi Xiong, Harry Perros, "Service Performance and Analysis in Cloud Computing," services, pp.693-700, *2009 Congress on Services - I*, 2009.

7.0 ACKNOWLEDGMENTS

The authors thank Michael Farber, Michael Cameron, and Paul Brown for funding this study through the Booz Allen Cloud Computing Initiative. The authors also thank Adam Young, Matt Ostrander, and Justin Neroda for their subject matter expertise and guidance.



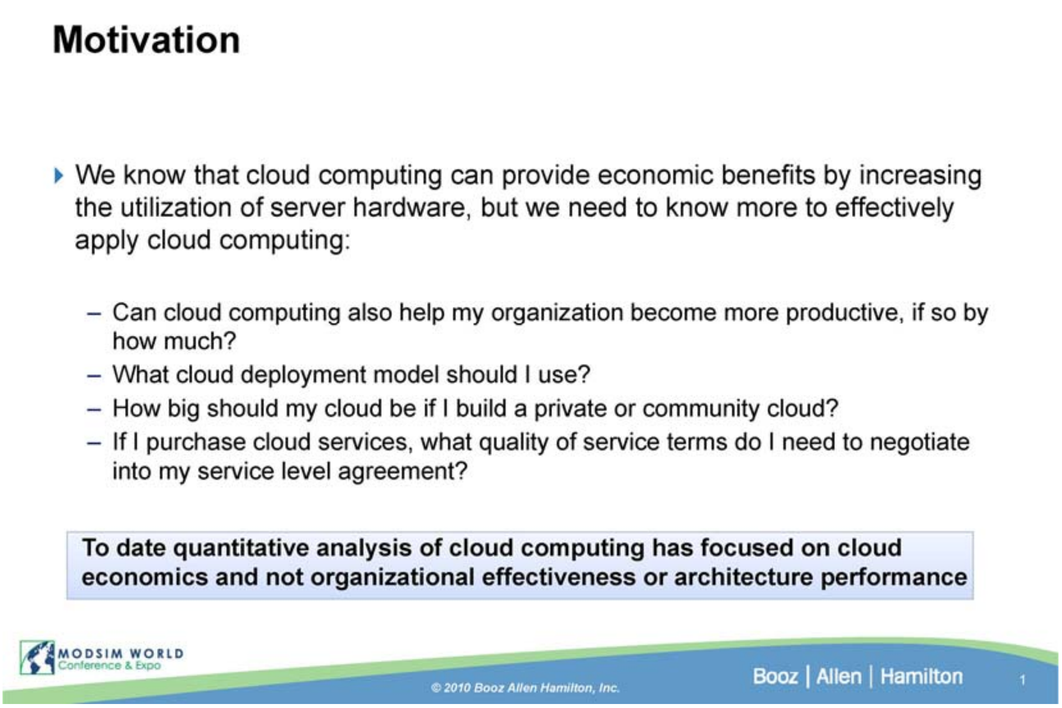
Booz | Allen | Hamilton

Performance Analysis of Cloud Computing Architectures using Discrete Event Simulation

John Stocker and Andrew Golomb

MODSIM WORLD Conference & Expo

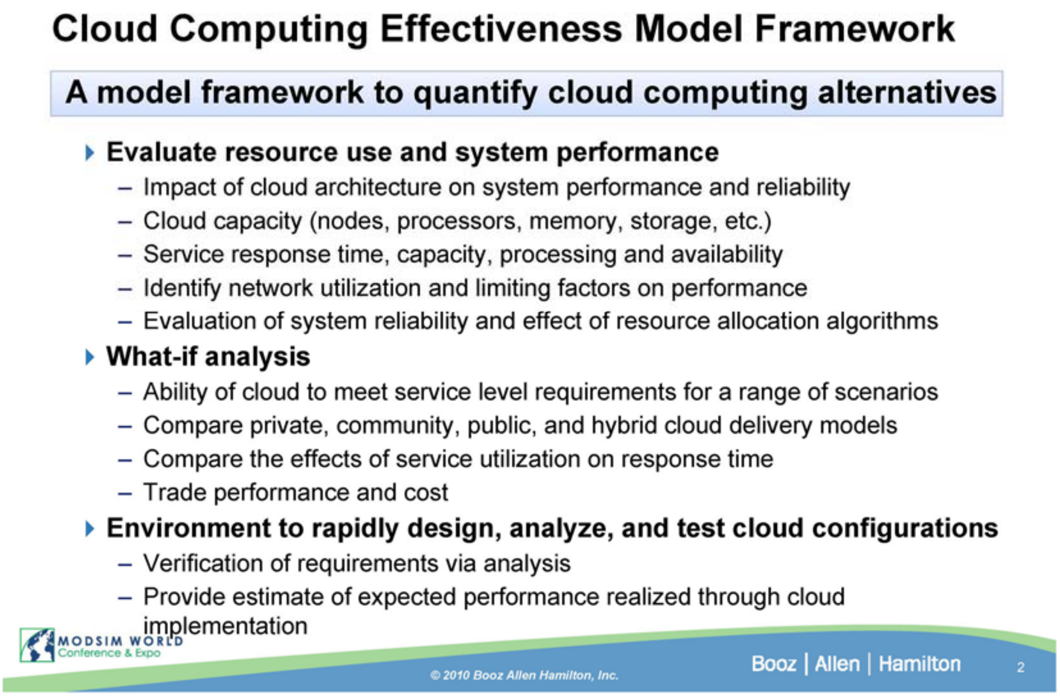
October 2010
Hampton, VA



Motivation

- We know that cloud computing can provide economic benefits by increasing the utilization of server hardware, but we need to know more to effectively apply cloud computing:
  - Can cloud computing also help my organization become more productive, if so by how much?
  - What cloud deployment model should I use?
  - How big should my cloud be if I build a private or community cloud?
  - If I purchase cloud services, what quality of service terms do I need to negotiate into my service level agreement?

To date quantitative analysis of cloud computing has focused on cloud economics and not organizational effectiveness or architecture performance



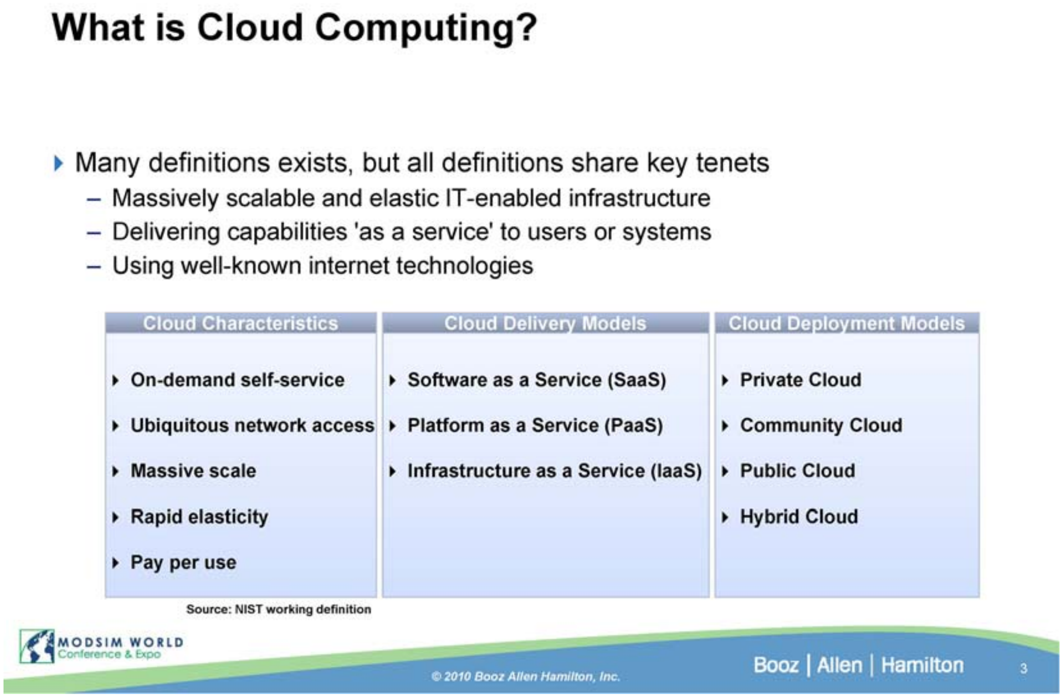
Cloud Computing Effectiveness Model Framework

A model framework to quantify cloud computing alternatives

- **Evaluate resource use and system performance**
  - Impact of cloud architecture on system performance and reliability
  - Cloud capacity (nodes, processors, memory, storage, etc.)
  - Service response time, capacity, processing and availability
  - Identify network utilization and limiting factors on performance
  - Evaluation of system reliability and effect of resource allocation algorithms
- **What-if analysis**
  - Ability of cloud to meet service level requirements for a range of scenarios
  - Compare private, community, public, and hybrid cloud delivery models
  - Compare the effects of service utilization on response time
  - Trade performance and cost
- **Environment to rapidly design, analyze, and test cloud configurations**
  - Verification of requirements via analysis
  - Provide estimate of expected performance realized through cloud implementation

MODSIM WORLD Conference & Expo

Booz | Allen | Hamilton 2



What is Cloud Computing?

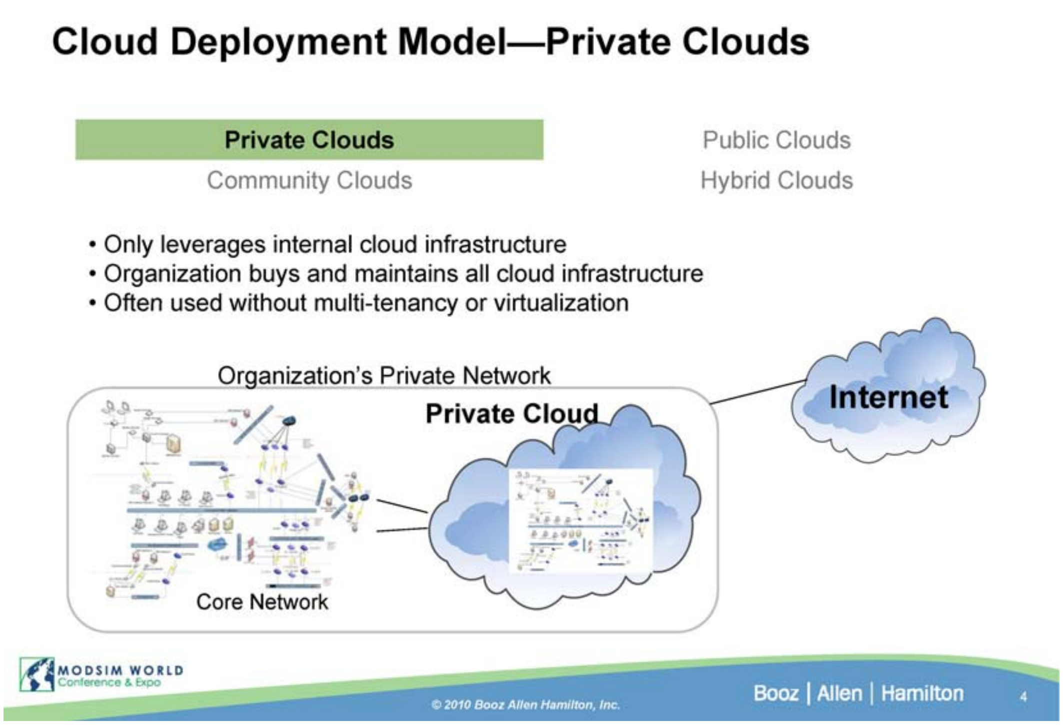
- Many definitions exists, but all definitions share key tenets
  - Massively scalable and elastic IT-enabled infrastructure
  - Delivering capabilities 'as a service' to users or systems
  - Using well-known internet technologies

Cloud Characteristics	Cloud Delivery Models	Cloud Deployment Models
On-demand self-service	Software as a Service (SaaS)	Private Cloud
Ubiquitous network access	Platform as a Service (PaaS)	Community Cloud
Massive scale	Infrastructure as a Service (IaaS)	Public Cloud
Rapid elasticity		Hybrid Cloud
Pay per use		

Source: NIST working definition

MODSIM WORLD Conference & Expo

Booz | Allen | Hamilton 3



Cloud Deployment Model—Private Clouds

Private Clouds | Public Clouds

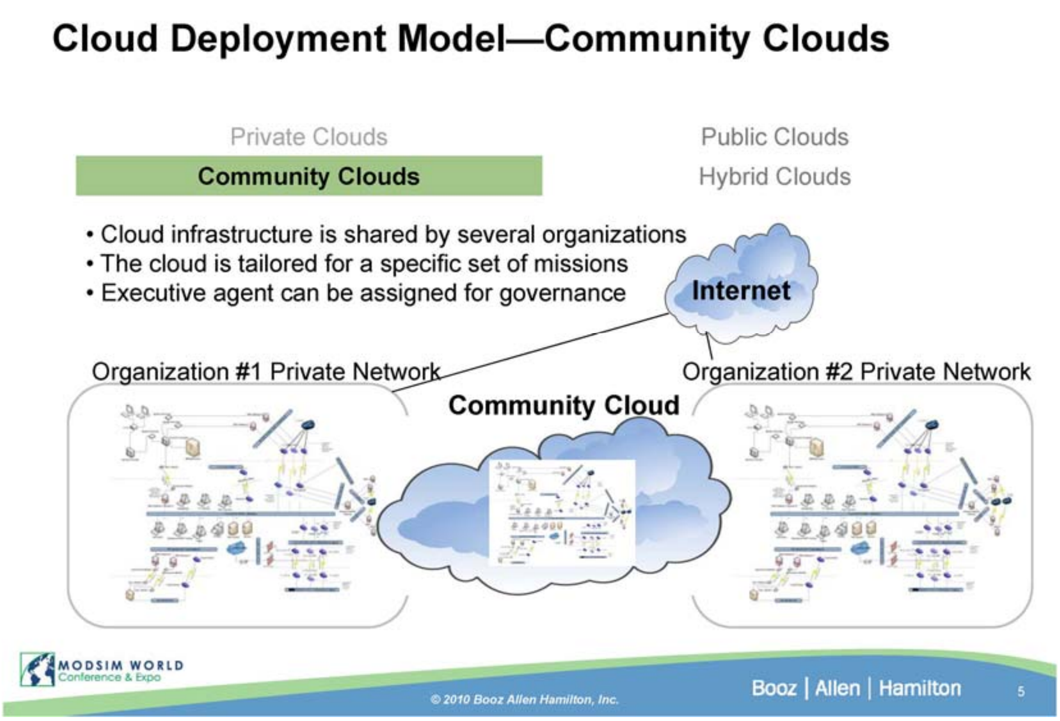
Community Clouds | Hybrid Clouds

- Only leverages internal cloud infrastructure
- Organization buys and maintains all cloud infrastructure
- Often used without multi-tenancy or virtualization

MODSIM WORLD Conference & Expo

© 2010 Booz Allen Hamilton, Inc.

Booz | Allen | Hamilton 4



Cloud Deployment Model—Community Clouds

Private Clouds | Public Clouds

Community Clouds | Hybrid Clouds

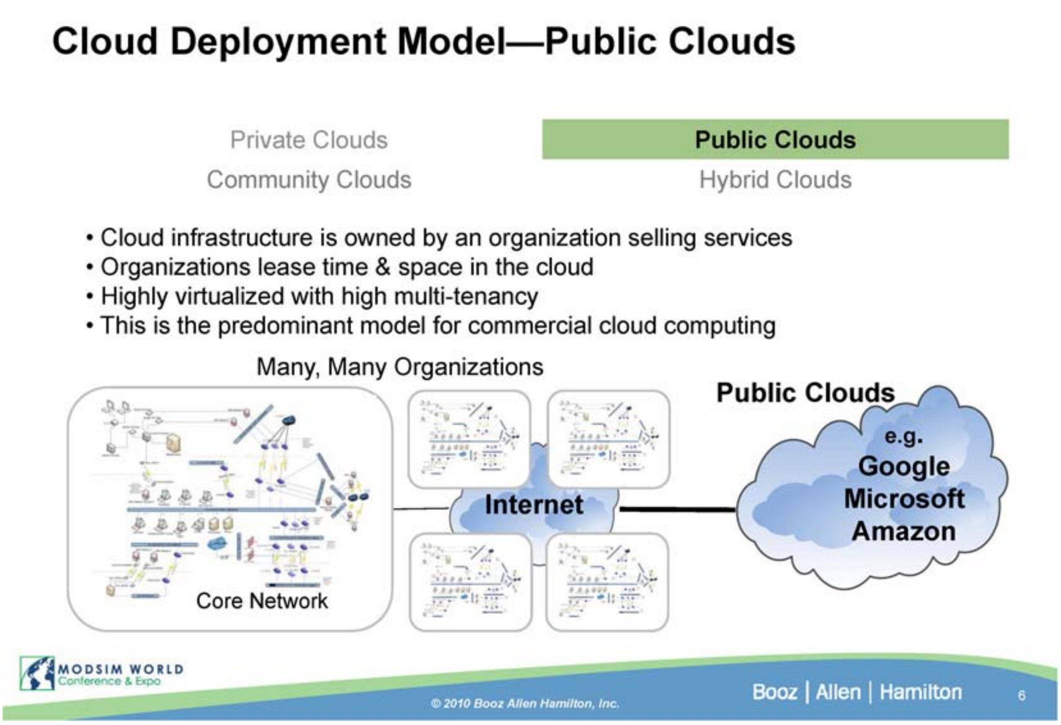
- Cloud infrastructure is shared by several organizations
- The cloud is tailored for a specific set of missions
- Executive agent can be assigned for governance

Internet
Organization #1 Private Network
Organization #2 Private Network
Community Cloud

MODSIM WORLD Conference & Expo

© 2010 Booz Allen Hamilton, Inc.

Booz | Allen | Hamilton 5



Cloud Deployment Model—Public Clouds

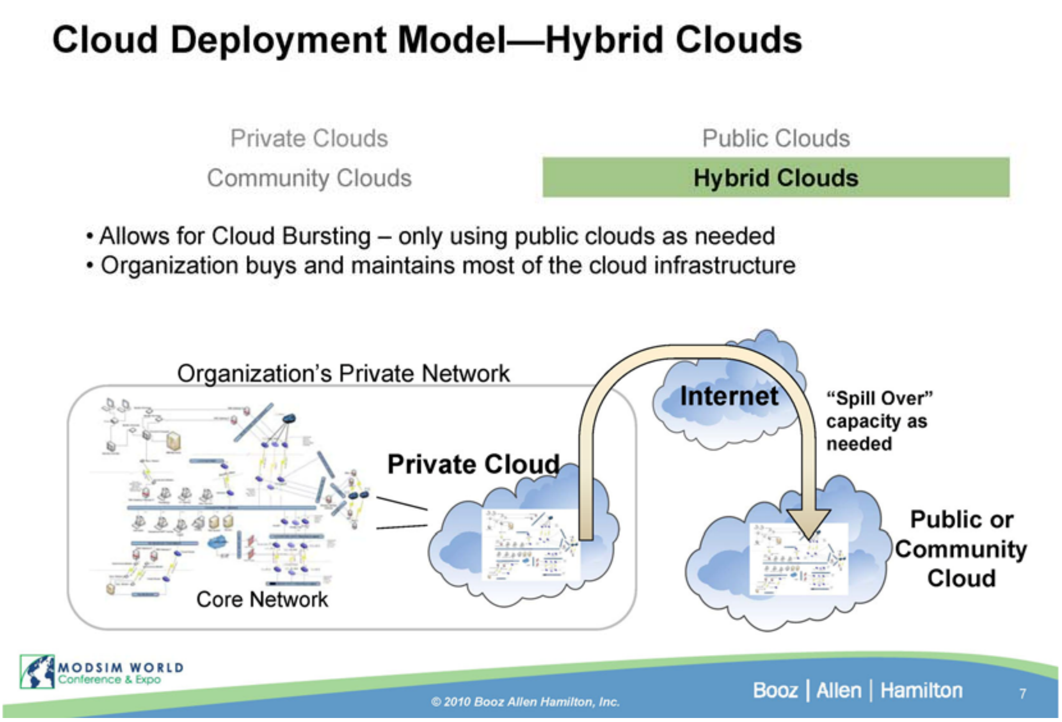
Private Clouds | **Public Clouds**

Community Clouds | Hybrid Clouds

- Cloud infrastructure is owned by an organization selling services
- Organizations lease time & space in the cloud
- Highly virtualized with high multi-tenancy
- This is the predominant model for commercial cloud computing

MODSIM WORLD Conference & Expo

Booz | Allen | Hamilton 6



Cloud Deployment Model—Hybrid Clouds

Private Clouds | Public Clouds

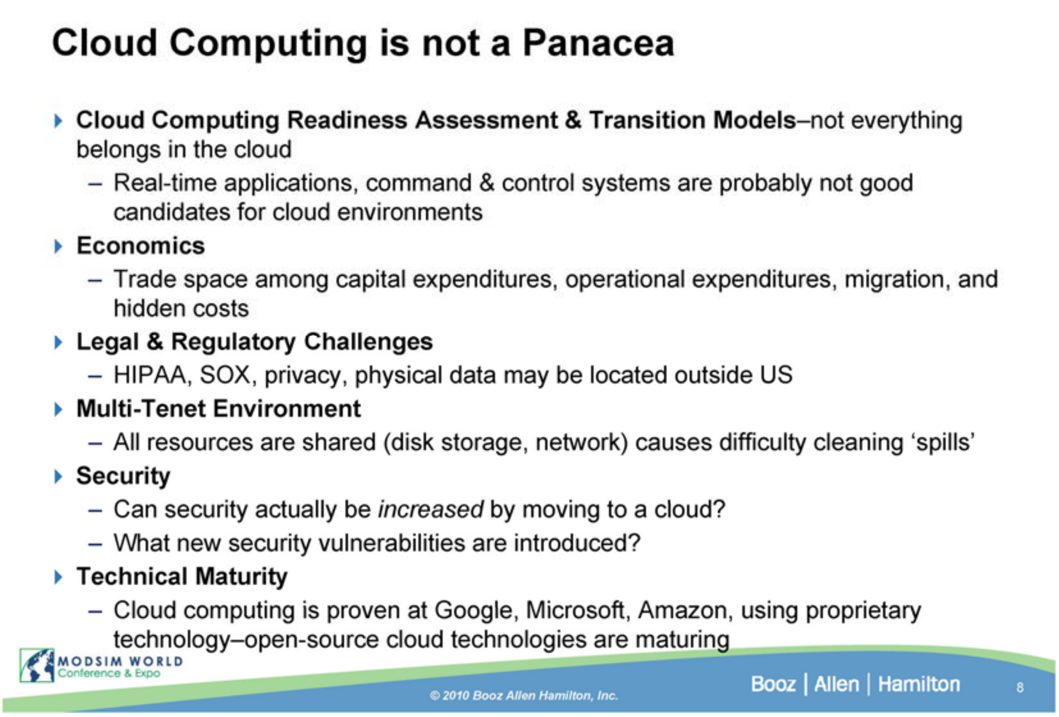
Community Clouds | **Hybrid Clouds**

- Allows for Cloud Bursting – only using public clouds as needed
- Organization buys and maintains most of the cloud infrastructure

Organization's Private Network
Core Network
Private Cloud
Internet
"Spill Over" capacity as needed
Public or Community Cloud

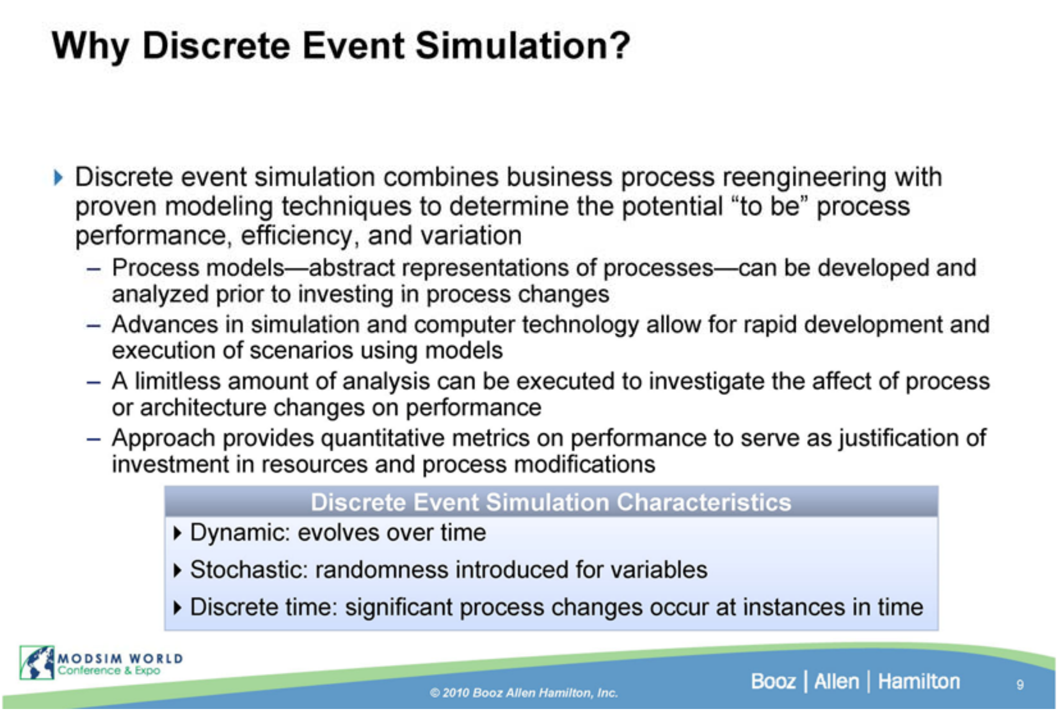
MODSIM WORLD Conference & Expo

Booz | Allen | Hamilton 7



Cloud Computing is not a Panacea

- **Cloud Computing Readiness Assessment & Transition Models**–not everything belongs in the cloud
  - Real-time applications, command & control systems are probably not good candidates for cloud environments
- **Economics**
  - Trade space among capital expenditures, operational expenditures, migration, and hidden costs
- **Legal & Regulatory Challenges**
  - HIPAA, SOX, privacy, physical data may be located outside US
- **Multi-Tenet Environment**
  - All resources are shared (disk storage, network) causes difficulty cleaning 'spills'
- **Security**
  - Can security actually be *increased* by moving to a cloud?
  - What new security vulnerabilities are introduced?
- **Technical Maturity**
  - Cloud computing is proven at Google, Microsoft, Amazon, using proprietary technology–open-source cloud technologies are maturing



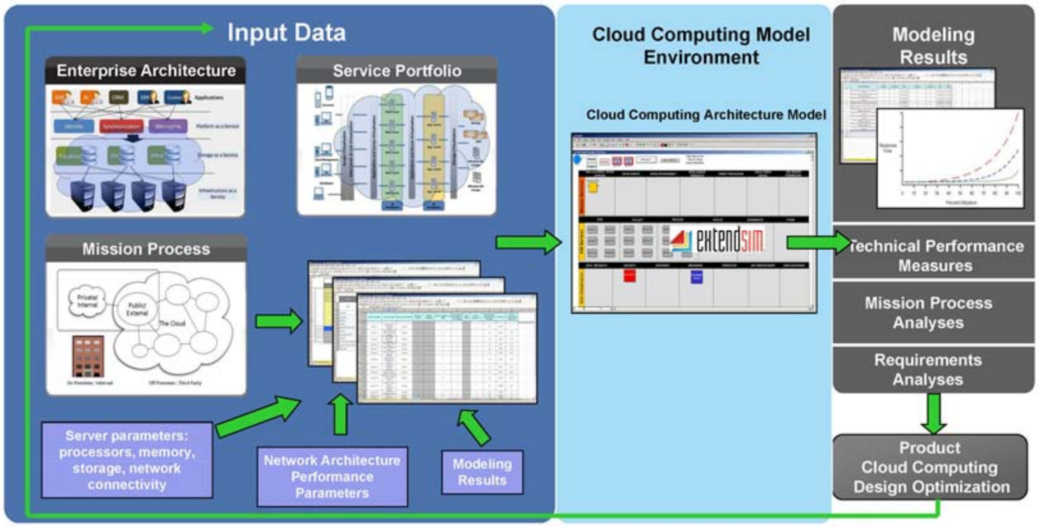
Why Discrete Event Simulation?

- Discrete event simulation combines business process reengineering with proven modeling techniques to determine the potential "to be" process performance, efficiency, and variation
  - Process models—abstract representations of processes—can be developed and analyzed prior to investing in process changes
  - Advances in simulation and computer technology allow for rapid development and execution of scenarios using models
  - A limitless amount of analysis can be executed to investigate the affect of process or architecture changes on performance
  - Approach provides quantitative metrics on performance to serve as justification of investment in resources and process modifications

Discrete Event Simulation Characteristics

- Dynamic: evolves over time
- Stochastic: randomness introduced for variables
- Discrete time: significant process changes occur at instances in time

Cloud Computing Performance Model



Input Data

Enterprise Architecture

Service Portfolio

Mission Process

Server parameters: processors, memory, storage, network connectivity

Network Architecture Performance Parameters

Modeling Results

Cloud Computing Model Environment

Cloud Computing Architecture Model

Modeling Results

Technical Performance Measures

Mission Process Analyses

Requirements Analyses

Product Cloud Computing Design Optimization

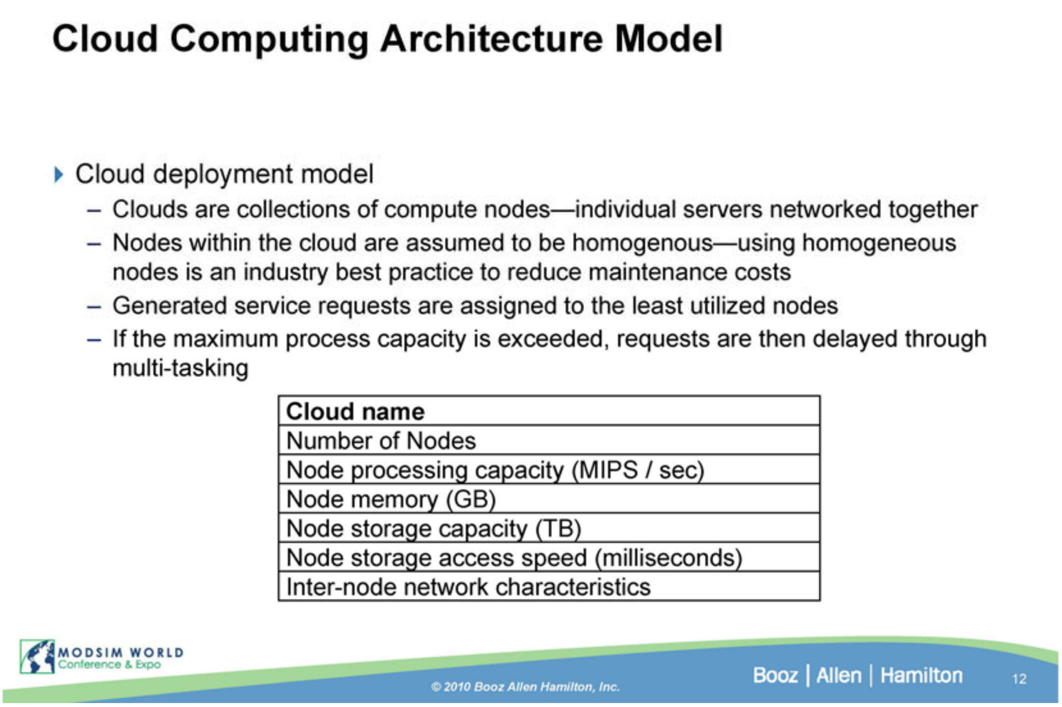
Business Process Model

Office site name
Number of staff members
Communication link characteristics - Maximum uplink bandwidth (Mbps) - Maximum downlink bandwidth (Mbps)

- Application and service model
  - Request rates vary over time: business day, weekly, monthly, quarterly, etc.
  - Computational requirements and data transferred are characterized by statistical distributions with application specific parameters

- Workforce model
  - The number of staff members per office site remains constant
  - Differences in application and service usage between staff members based on their roles is aggregated in the application request rate
  - Uplink and downlink bandwidth between the office and the cloud

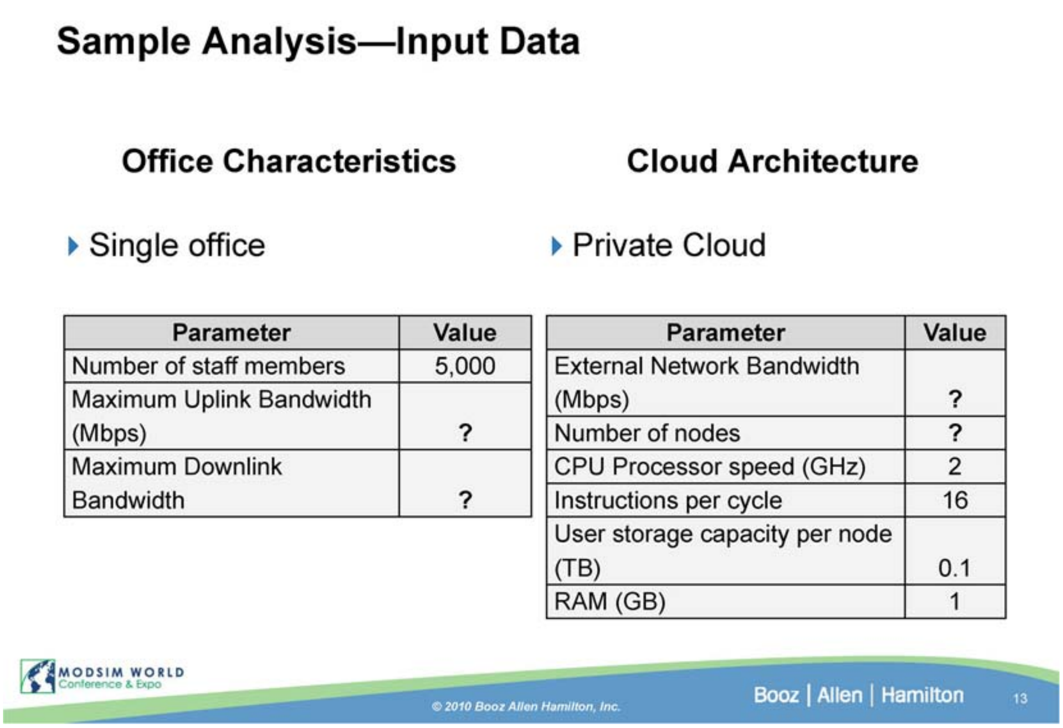
Application or service name
Request rate (requests over time)
Data transferred (KB)
Computational requirements (MIPS)
QoS requirements - Allowable response latency (seconds)



Cloud Computing Architecture Model

- Cloud deployment model
  - Clouds are collections of compute nodes—individual servers networked together
  - Nodes within the cloud are assumed to be homogenous—using homogeneous nodes is an industry best practice to reduce maintenance costs
  - Generated service requests are assigned to the least utilized nodes
  - If the maximum process capacity is exceeded, requests are then delayed through multi-tasking

Cloud name
Number of Nodes
Node processing capacity (MIPS / sec)
Node memory (GB)
Node storage capacity (TB)
Node storage access speed (milliseconds)
Inter-node network characteristics



Sample Analysis—Input Data

Office Characteristics

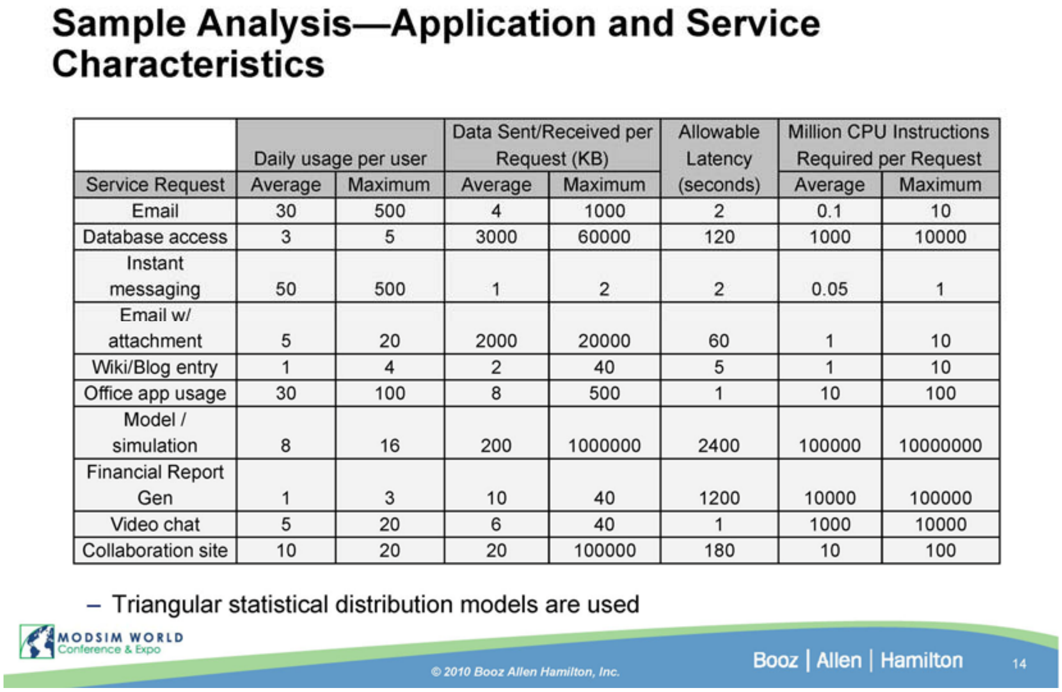
- Single office

Parameter	Value
Number of staff members	5,000
Maximum Uplink Bandwidth (Mbps)	**?**
Maximum Downlink Bandwidth	**?**

Cloud Architecture

- Private Cloud

Parameter	Value
External Network Bandwidth (Mbps)	**?**
Number of nodes	**?**
CPU Processor speed (GHz)	2
Instructions per cycle	16
User storage capacity per node (TB)	0.1
RAM (GB)	1



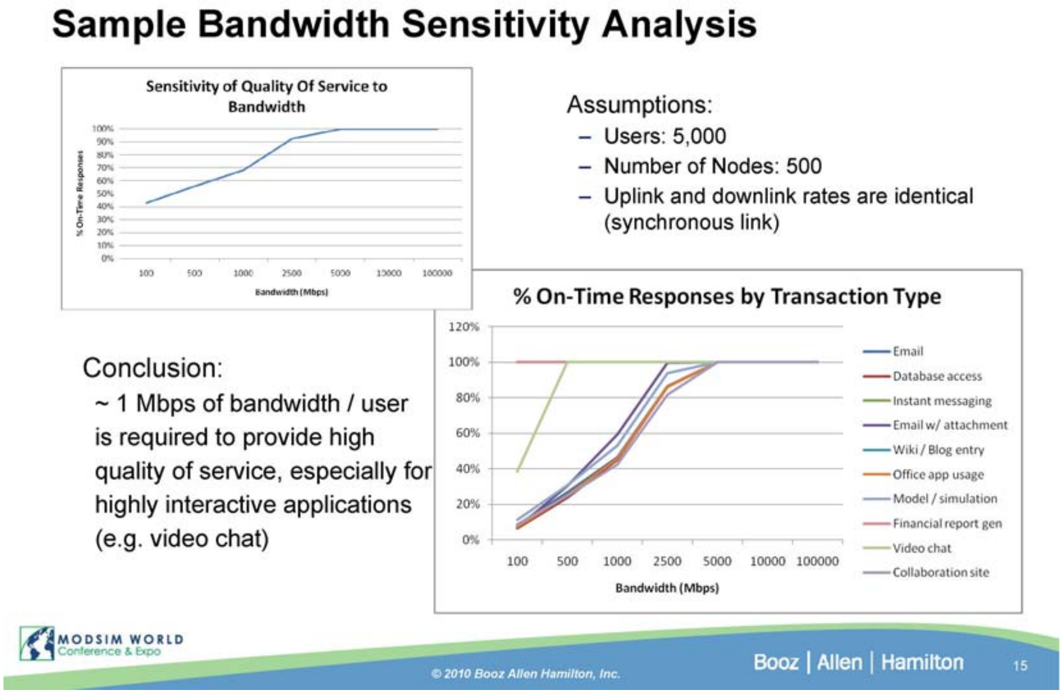
Sample Analysis—Application and Service Characteristics

	Daily usage per user		Data Sent/Received per Request (KB)		Allowable Latency	Million CPU Instructions Required per Request	
Service Request	Average	Maximum	Average	Maximum	(seconds)	Average	Maximum
Email	30	500	4	1000	2	0.1	10
Database access	3	5	3000	60000	120	1000	10000
Instant messaging	50	500	1	2	2	0.05	1
Email w/ attachment	5	20	2000	20000	60	1	10
Wiki/Blog entry	1	4	2	40	5	1	10
Office app usage	30	100	8	500	1	10	100
Model / simulation	8	16	200	1000000	2400	100000	10000000
Financial Report Gen	1	3	10	40	1200	10000	100000
Video chat	5	20	6	40	1	1000	10000
Collaboration site	10	20	20	100000	180	10	100

– Triangular statistical distribution models are used

MODSIM WORLD Conference & Expo

Booz | Allen | Hamilton



Sample Bandwidth Sensitivity Analysis

Assumptions:

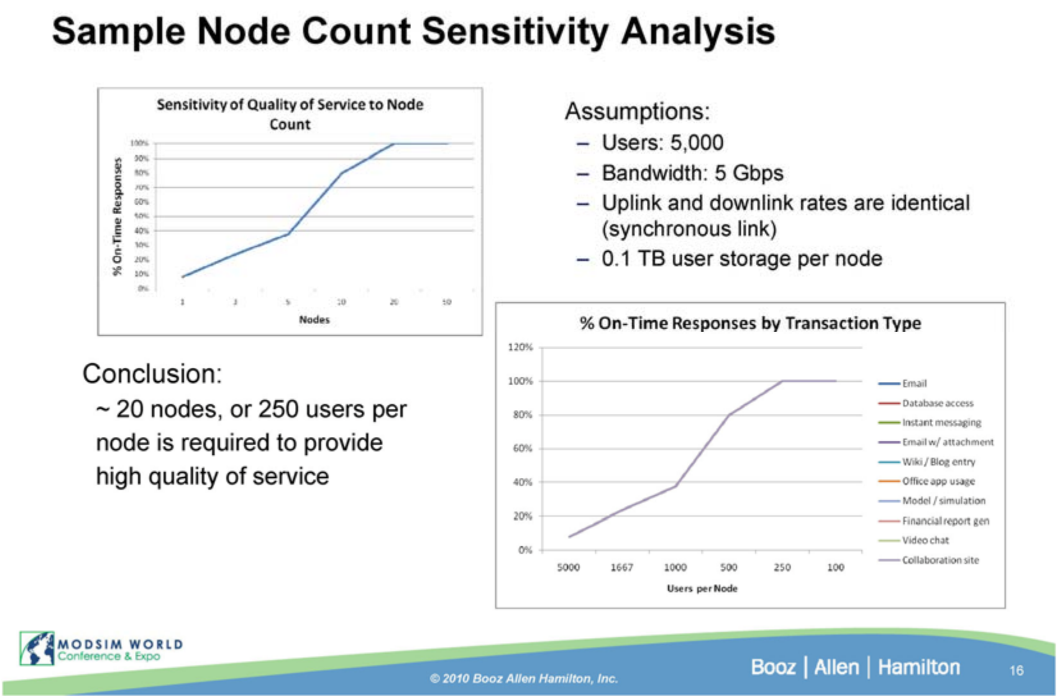
- Users: 5,000
- Number of Nodes: 500
- Uplink and downlink rates are identical (synchronous link)

Conclusion:

~ 1 Mbps of bandwidth / user is required to provide high quality of service, especially for highly interactive applications (e.g. video chat)

MODSIM WORLD Conference & Expo

Booz | Allen | Hamilton



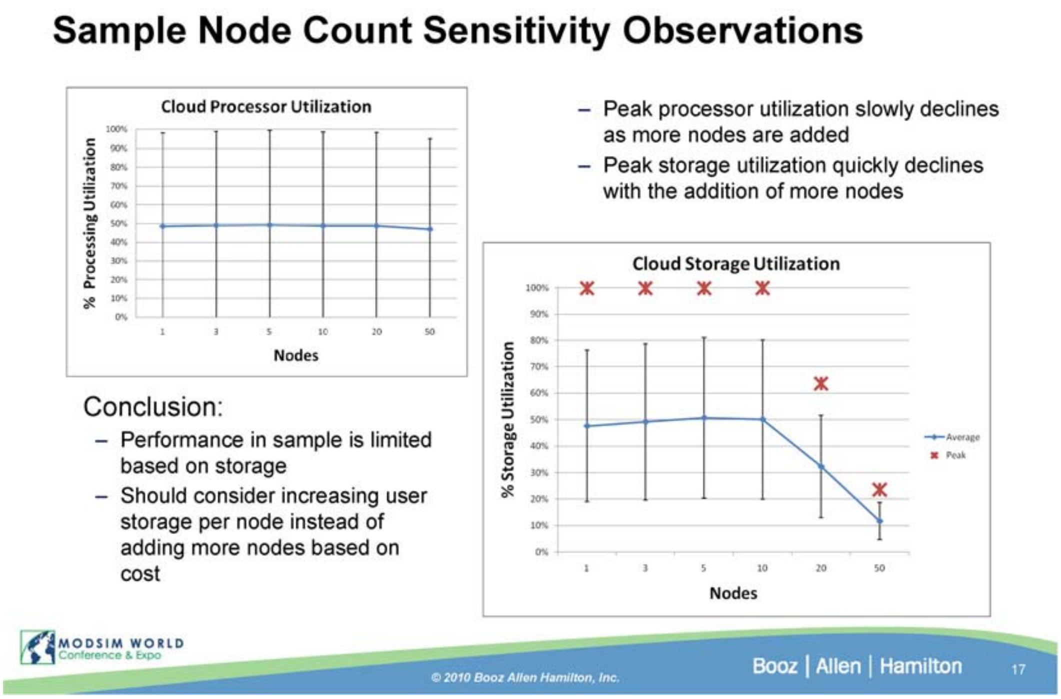
Sample Node Count Sensitivity Analysis

Assumptions:

- Users: 5,000
- Bandwidth: 5 Gbps
- Uplink and downlink rates are identical (synchronous link)
- 0.1 TB user storage per node

Conclusion:

~ 20 nodes, or 250 users per node is required to provide high quality of service



Sample Node Count Sensitivity Observations

- Peak processor utilization slowly declines as more nodes are added
- Peak storage utilization quickly declines with the addition of more nodes

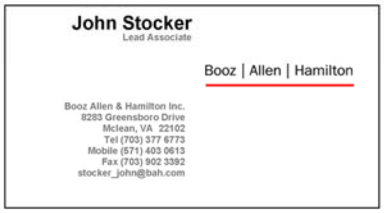
Conclusion:

- Performance in sample is limited based on storage
- Should consider increasing user storage per node instead of adding more nodes based on cost

Conclusion and Future Work

- We have developed a model framework for analyzing the performance of candidate cloud computing architectures and resultant organizational effectiveness
- Baseline future models using data from Booz Allen's cloud computing infrastructure that supports 30,000+ employees
- Business process modeling can be improved through integrating industry standard process model data captured in FEAF or DoDAF artifacts
- Coupling quantitative performance and a cost models will provide an overarching analytical approach to develop cloud migration strategies
- Detailed modeling and analysis of cloud computing architectures is necessary to capture trades among:
  - Cloud deployment models: private, community, public, hybrid
  - Fewer more powerful nodes vs. many less powerful nodes
  - Commodity hardware vs. more expensive, high-availability hardware

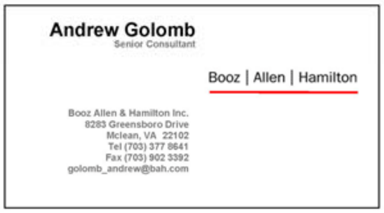
Questions and Comments



John Stocker
Lead Associate

Booz | Allen | Hamilton

Booz Allen & Hamilton Inc.
8283 Greensboro Drive
Mclean, VA 22102
Tel (703) 377 6773
Mobile (571) 403 0613
Fax (703) 902 3392
stocker_john@bah.com

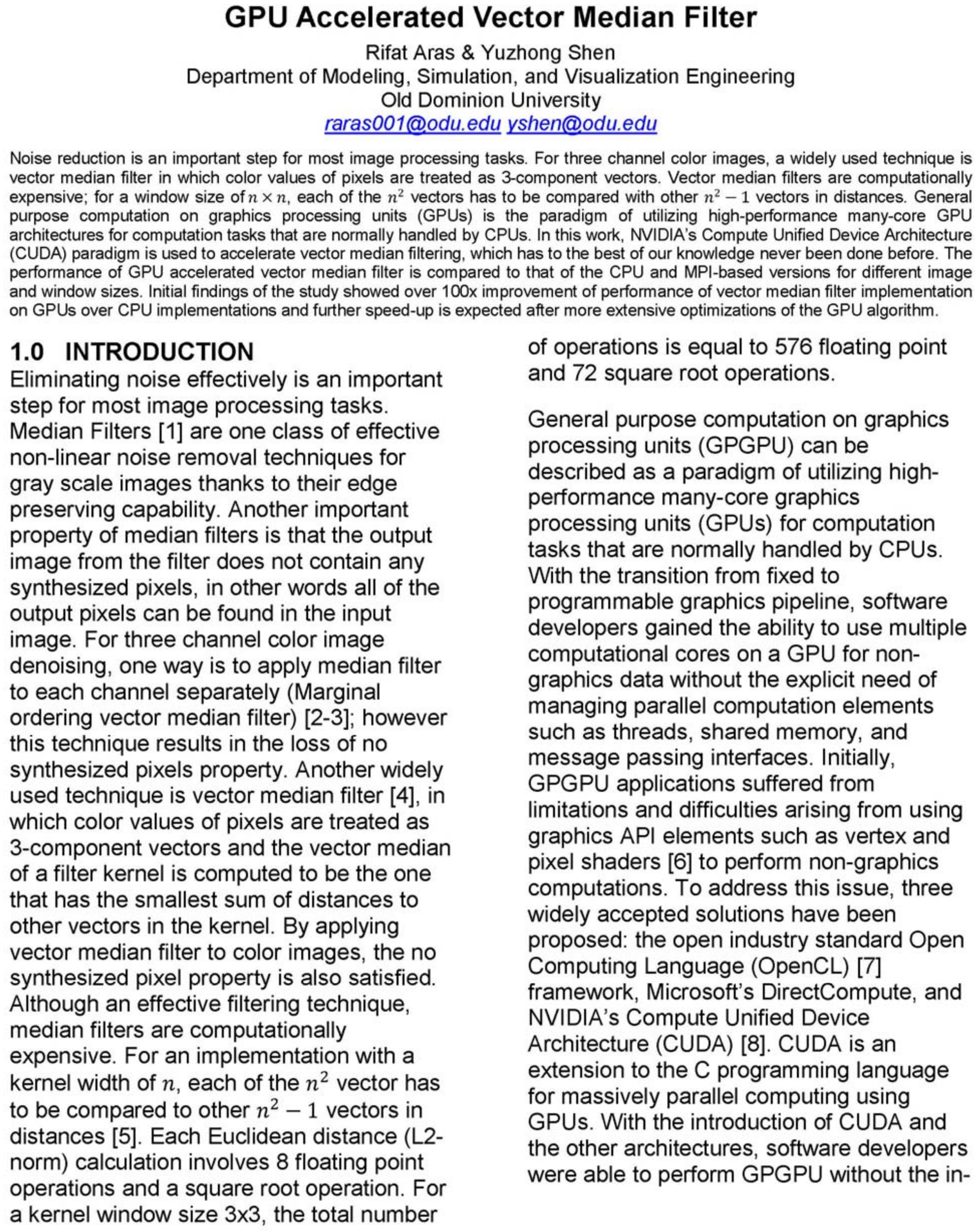


Andrew Golomb
Senior Consultant

Booz | Allen | Hamilton

Booz Allen & Hamilton Inc.
8283 Greensboro Drive
Mclean, VA 22102
Tel (703) 377 8641
Fax (703) 902 3392
golomb_andrew@bah.com

2.4 GPU Accelerated Vector Median Filter



GPU Accelerated Vector Median Filter

Rifat Aras & Yuzhong Shen
Department of Modeling, Simulation, and Visualization Engineering
Old Dominion University
raras001@odu.edu yshen@odu.edu

Noise reduction is an important step for most image processing tasks. For three channel color images, a widely used technique is vector median filter in which color values of pixels are treated as 3-component vectors. Vector median filters are computationally expensive; for a window size of $n \times n$, each of the n^2 vectors has to be compared with other $n^2 - 1$ vectors in distances. General purpose computation on graphics processing units (GPUs) is the paradigm of utilizing high-performance many-core GPU architectures for computation tasks that are normally handled by CPUs. In this work, NVIDIA's Compute Unified Device Architecture (CUDA) paradigm is used to accelerate vector median filtering, which has to the best of our knowledge never been done before. The performance of GPU accelerated vector median filter is compared to that of the CPU and MPI-based versions for different image and window sizes. Initial findings of the study showed over 100x improvement of performance of vector median filter implementation on GPUs over CPU implementations and further speed-up is expected after more extensive optimizations of the GPU algorithm.

1.0 INTRODUCTION

Eliminating noise effectively is an important step for most image processing tasks. Median Filters [1] are one class of effective non-linear noise removal techniques for gray scale images thanks to their edge preserving capability. Another important property of median filters is that the output image from the filter does not contain any synthesized pixels, in other words all of the output pixels can be found in the input image. For three channel color image denoising, one way is to apply median filter to each channel separately (Marginal ordering vector median filter) [2-3]; however this technique results in the loss of no synthesized pixels property. Another widely used technique is vector median filter [4], in which color values of pixels are treated as 3-component vectors and the vector median of a filter kernel is computed to be the one that has the smallest sum of distances to other vectors in the kernel. By applying vector median filter to color images, the no synthesized pixel property is also satisfied. Although an effective filtering technique, median filters are computationally expensive. For an implementation with a kernel width of n, each of the n^2 vector has to be compared to other $n^2 - 1$ vectors in distances [5]. Each Euclidean distance (L2-norm) calculation involves 8 floating point operations and a square root operation. For a kernel window size 3x3, the total number of operations is equal to 576 floating point and 72 square root operations.

General purpose computation on graphics processing units (GPGPU) can be described as a paradigm of utilizing high-performance many-core graphics processing units (GPUs) for computation tasks that are normally handled by CPUs. With the transition from fixed to programmable graphics pipeline, software developers gained the ability to use multiple computational cores on a GPU for non-graphics data without the explicit need of managing parallel computation elements such as threads, shared memory, and message passing interfaces. Initially, GPGPU applications suffered from limitations and difficulties arising from using graphics API elements such as vertex and pixel shaders [6] to perform non-graphics computations. To address this issue, three widely accepted solutions have been proposed: the open industry standard Open Computing Language (OpenCL) [7] framework, Microsoft's DirectCompute, and NVIDIA's Compute Unified Device Architecture (CUDA) [8]. CUDA is an extension to the C programming language for massively parallel computing using GPUs. With the introduction of CUDA and the other architectures, software developers were able to perform GPGPU without the in-depth knowledge of programmable graphics shaders.

In this work, our main contribution is implementing vector median filter using the CUDA programming paradigm and applying CUDA specific optimizations. The performance of the implemented filter is compared to the single thread CPU and multi-processor MPI versions with respect to different image and kernel sizes.

The remainder of the paper is organized as follows. Section 2 describes the definition of vector median filter, previous attempts for accelerating vector median filters, and CUDA and MPI implementations. Section 3 compares the different implementations of the filter in terms of performance. Finally, Section 4 concludes the paper and discusses future work.

2.0 BODY

2.1 Vector Median Filters

Non-linear filters such as Bilateral [9] and Median filters are important image processing techniques of gray scale and colored image processing because of their ability to preserve edge, line, and other image structures while removing noise artifacts. Vector median filter performs non-linear filtering by moving a window over a pixel (Fig. 1) (with RGB channels) and selects the pixel that has the smallest sum of distance to the other pixels in the window as the output [4].

Given a window that contains $N = n \times n$ pixels denoted by $W = \{x_1, x_2, \dots, x_N\}$ the output of vector median filter x_{VM} is computed by Eq. (1).

$$\sum_{i=1}^{N} \|x_{VM} - x_i\| \leq \sum_{i=1}^{N} \|x_j - x_i\|, j = 1, \dots, N, \quad (1)$$

where $\|\cdot\|$ denotes the distance metric between the vectors. In this paper Euclidean distance (L2-norm) is used to determine the distance between two vectors. The Euclidean distance between two vectors u and v that are in $\mathbb{R}^p$ is given in Eq. 2.

$$\|u - v\| = \sqrt{\sum_{k=1}^{p} (u_{(k)} - v_{(k)})^2}. \quad (2)$$

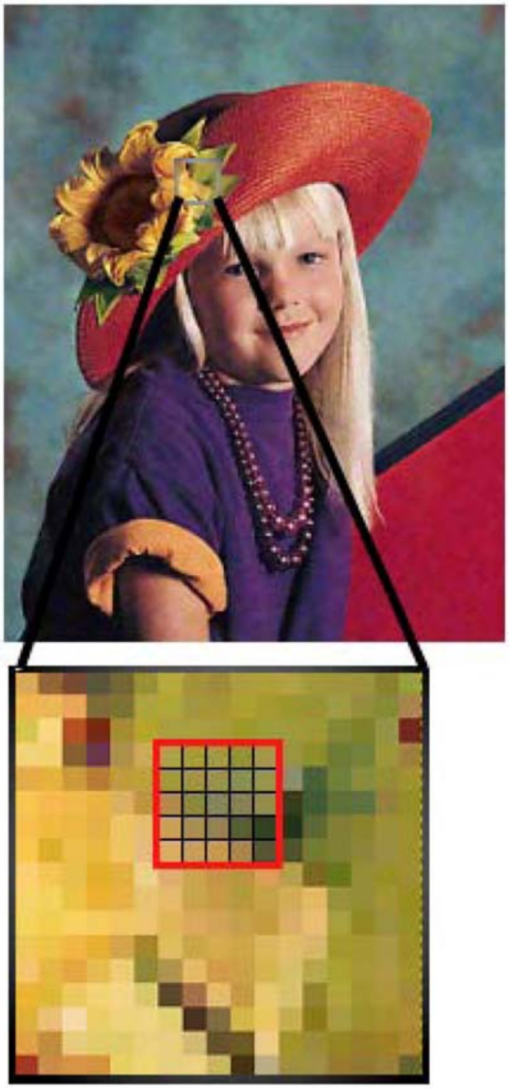


Figure 1. An example 5x5 window. This window slides over the target pixels and compute the output according to the 25 pixel values inside.

Variations of vector median filters have been developed to address wide variety of problem domain. Weighted vector median filters [10] fuzzy vector median filters [11] are two variations of vector median filter that have been successfully deployed in a number of applications.

The computational complexity of vector median filter makes it very challenging to be used for large problems that have stringent time requirements. To address this issue, there have been numerous attempts to accelerate the vector median computation by different means.

In the work of Boudabous et al. [12], a parallel architecture of the vector median filter was implemented in an embedded system. The time consuming steps of the filter were implemented in hardware level, specifically by programming Field Programmable Gate Array (FPGA) devices with VHSIC Hardware Description Language (VHDL).

Another approach to accelerate the vector median filter was proposed by Barni and Cappellini [13]. In this work, the performance of the filter is increased by using the L1-norm distance metric instead of the L2-norm. In addition to using the simplified distance metric, another simplification the authors adopted is using a central color for comparisons. In other words, for a window of pixels, the output pixel is chosen to be the one that is closest to the central color that is obtained by component wise application (marginal) of median filter to the color channels. These simplifications increase the performance of the filter significantly, however at a cost of decreased quality of the output.

There are other attempts to accelerate vector median filters by utilizing different distance metrics. For a list of such work and their computational complexity, the reader is encouraged to refer to the work of Barni and Cappellini [5].

2.2 A Brief CUDA Primer

For more than two decades, the end users of computer systems with single central processing units (CPU) enjoyed the increasing performance of their applications with each new generation of CPUs. The equation was simple, as the clock frequency of the CPUs increased with each generation; the very same application was able to run faster on the new architecture. However, this profile of increasing speeds has slowed down in 2003 due to the power consumption and heat dissipation issues [14]. The responses of CPU manufacturers to address these limitations were producing multi-core CPUs having similar clock frequencies with previous generations. With this adopted strategy, the expectation of increased performance with new generation of CPUs vanished especially for the so called sequential applications that rely on a single CPU. In order to satisfy the performance demand of the end users, application developers need to delve into the art of parallel programming, which is typically performed on large-scale expensive parallel computers, such as clusters.

Since 2003, microprocessor manufacturers adopted two main strategies for their processor designs. The multi-core strategy (CPU designs) provides small number of large cores (typically 2-4 cores) that try to maximize the execution speed of sequential programs with their large control logic elements and cache structures. On the other hand, designs that adopt many-core strategy (GPU designs) try to maximize the floating point calculation throughput by devoting more processor area (typically 120-240 cores) to data processing units instead of flow control logic elements and large data caches [8]. Because of these design choices, the recent ratio of theoretical computation peak of GPUs over CPUs is roughly 6.5 to 1 (1300 Gflop/s to 200 Gflop/s). When the huge difference between the two architecture's memory bandwidth is included in the equation (Fig. 2), GPUs turn out to be well suited for massively parallel applications with high arithmetic intensity, in which the same instruction is applied to multiple data [8].

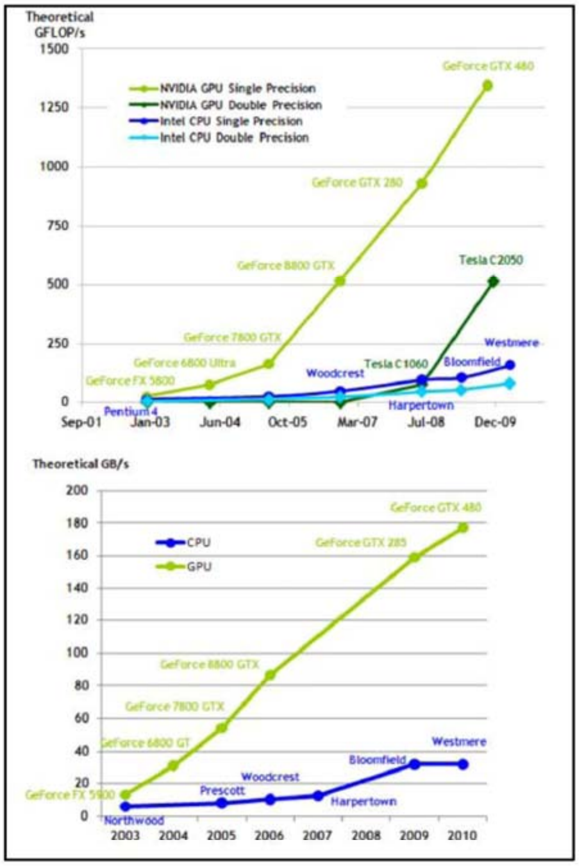


Figure 2. The comparison of GPU and CPU architectures in terms of computation power and memory bandwidth [8].

General-purpose programming using graphics processing units (GPGPU) became an active research area that studies the methods and algorithms for a wide range of problem domain such as signal and image processing, physically based simulations, computational finance, and computational biology [15-16]. The first generation of GPGPU applications were difficult to implement because programmers had to use interfaces that were primarily designed for computer graphics computations such as C for graphics (Cg) by NVIDIA, High Level Shading Language (HLSL) by Microsoft, and OpenGL Shading Language (GLSL) by Khronos Group. These computer graphics oriented APIs limited the kinds of applications that can work on GPUs [14].

To address this issue, in 2007 NVIDIA released Compute Unified Device Architecture (CUDA), which is not only an extension to the C programming language, but also adding additional hardware to the chip to facilitate the ease of parallel programming [14]. In the CUDA enabled chips, CUDA programs do not go through the graphics interface. CUDA requests are handled by a new general-purpose parallel programming interface located on the chip. CUDA relieves GPGPU application developers the necessity of having in-depth knowledge of graphics interfaces or programmable shaders.

CUDA also has significant advantages over classical parallel programming languages and models such as Message Passing Interface (MPI) for scalable cluster computing and OpenMP for shared-memory multiprocessor systems. MPI is for cluster systems, in which data sharing is done by explicit message passing, in other words processors in a cluster do not share memory. Parallel applications written using MPI model have been known to run on clusters with more than 100,000 processors. OpenMP, on the other hand, can support shared memory interface. However, it suffers from scalability issues. Parallel applications using OpenMP could not be able to scale beyond a couple hundred computing nodes mainly because of the thread management overheads [14]. Compared to these legacy models, CUDA provides a shared memory interface among the cores of a streaming multiprocessor (SM) along with higher scalability and low-overhead thread management properties.

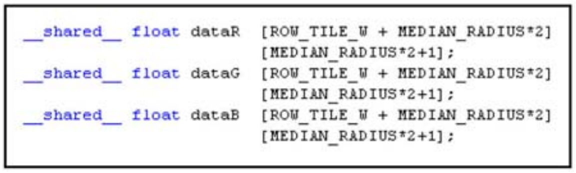
CUDA enabled GPU architectures consist of arrays of streaming multiprocessors (SM). Different generations of GPUs contain different number of heavily threaded SMs. Each SM contains a total of 8 streaming processors (SPs) or in other words cores. Cores of the same SM share the control logic, instruction cache and fast access shared memory. GPUs support up to 4 GBs of graphics double data rate (GDDR) DRAM that serves as the frame buffer and texture memory for 3D graphics applications. For general purpose computations, this memory space is referred as the global memory that resides off-chip and has very high bandwidth. Each SP has a multiply-add (MAD) unit, an additional multiply unit, and special function units performing floating point operations such as square roots. Because of massively threaded nature of SPs, thousands of concurrent threads can be handled for a GPGPU application. The recent GPUs with GT200 architecture can support 1024 threads per SM, which roughly sums up to about 30,000 threads for the entire chip [8].

In CUDA, fine-grained data parallelism is achieved by the massively threaded structure. Kernels are user created functions that contain statements that are executed by each individual thread. These threads are organized into a two-level hierarchy: 3D blocks consisting of individual threads and 2D grid consisting of blocks. The threads in the blocks are further divided into groups of 32 threads called warps. The notion of warp is important, because a warp is the unit of thread scheduling in SMs, i.e. when a warp is scheduled to run, all of the accompanying threads run the same single instruction. CUDA also scales transparently with the underlying hardware capabilities. Without changing the underlying code, the GPGPU application can run on different GPU hardware that may have different number of SMs or thread capacities. This important property of CUDA programs is achieved by allowing the execution of blocks in any order [8]. When a kernel is launched, the threads of the kernel are distributed among SMs on a block by block basis. Each SM can claim at most 8 blocks at a time. When more blocks are involved than the maximum number of resident blocks (# of SMs x 8) in a CUDA application, the maintained list of blocks that need to be executed is used and new blocks are assigned to SMs as they complete the execution of previously assigned blocks.

In order to obtain the maximum performance out of the CUDA capable GPUs, the memory hierarchy of CUDA has to be examined carefully. Global memory and constant memory are located at the bottom. Global memory (or device memory) supports high-bandwidth read-write access that has relatively higher access latency compared to system's RAM. Because of this relatively higher latency, it should be used wisely not to cause performance degradation. Constant memory supports short-latency high-bandwidth read-only cached access by the device. Shared memory is an on-chip memory that is allocated to thread blocks. Shared memory is a very fast parallel access enabled memory space, which is often used by the threads of the same block to cooperate by sharing their input data and intermediate results. Registers are at the top of the hierarchy. They are allocated privately for each thread and typically used to hold frequently accessed variables other than arrays. Besides these basic memory structures, CUDA also provides a texture memory space that resides in global memory but cached in texture cache. The texture cache makes use of 2D spatial locality, i.e. if threads of the same warp access texture addresses that are close in 2D space, they can achieve fast access rates [8, 14]. Special care has to be taken when accessing these memory spaces. Because of the underlying DRAM structure of global memory, it is used most efficiently when threads of the same warp access it in a certain pattern. Shared memory accesses also require caution for fast access. To support parallel access, shared memory space is divided into equally sized memory banks that can be accessed simultaneously as long as different threads in a warp access different shared memory banks. If different threads try to access memory addresses that reside in the same bank, a bank conflict occurs and the accesses are serialized. More information about memory access patterns can be found in NVIDIA CUDA C Programming Guide [8].

2.3 CUDA Implementation

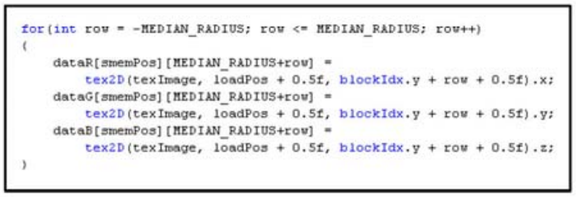
To utilize the massively threaded nature of CUDA, we followed a divide and conquer approach on the input image. The image is divided into tiles of size 1 row and N columns that are assigned to thread blocks for processing (Fig. 6). Each thread therefore is responsible for accessing global memory, copying the pixel data to shared memory space (Fig. 3), and computing the median filter output for a single pixel.



```
__shared__ float dataR  [ROW_TILE_W + MEDIAN_RADIUS*2]
                        [MEDIAN_RADIUS*2+1];
__shared__ float dataG  [ROW_TILE_W + MEDIAN_RADIUS*2]
                        [MEDIAN_RADIUS*2+1];
__shared__ float dataB  [ROW_TILE_W + MEDIAN_RADIUS*2]
                        [MEDIAN_RADIUS*2+1];
```

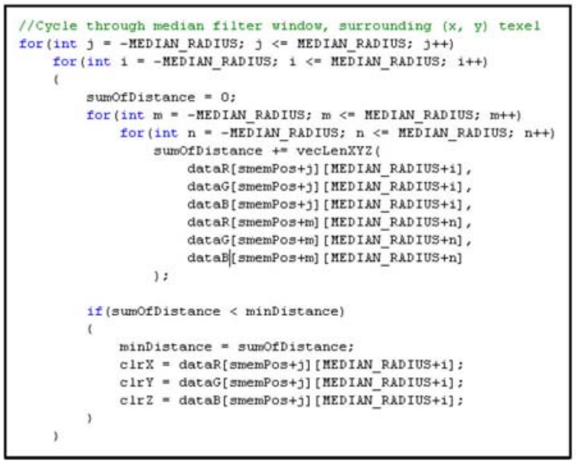
Figure 3. Arrays located in shared memory space that store color information of pixels. The color channels are stored separately in order to access the shared memory without bank conflicts.

As median filter works in a window of pixels, individual threads are also responsible for accessing and copying the neighboring pixels. After the pixel data is copied to the shared memory by following the coalesced global memory access requirements (Fig. 4), each thread computes the median vector among the vectors in the surrounding window. The median vector computation is comprised of 4 nested loops (Fig. 5) and takes the major part of the running time. When the median vector is computed, it is written back to the global memory, which is finally read back to the CPU.



```
for(int row = -MEDIAN_RADIUS; row <= MEDIAN_RADIUS; row++)
{
    dataR[smemPos][MEDIAN_RADIUS+row] =
        tex2D(texImage, loadPos + 0.5f, blockIdx.y + row + 0.5f).x;
    dataG[smemPos][MEDIAN_RADIUS+row] =
        tex2D(texImage, loadPos + 0.5f, blockIdx.y + row + 0.5f).y;
    dataB[smemPos][MEDIAN_RADIUS+row] =
        tex2D(texImage, loadPos + 0.5f, blockIdx.y + row + 0.5f).z;
}
```

Figure 4. Loading the shared memory arrays with pixel color data. The pixels that are in MEDIAN_RADIUS neighborhood are also loaded.



```
//Cycle through median filter window, surrounding (x, y) texel
for(int j = -MEDIAN_RADIUS; j <= MEDIAN_RADIUS; j++)
    for(int i = -MEDIAN_RADIUS; i <= MEDIAN_RADIUS; i++)
    {
        sumOfDistance = 0;
        for(int m = -MEDIAN_RADIUS; m <= MEDIAN_RADIUS; m++)
            for(int n = -MEDIAN_RADIUS; n <= MEDIAN_RADIUS; n++)
                sumOfDistance += vecLenXYZ(
                    dataR[smemPos+j][MEDIAN_RADIUS+i],
                    dataG[smemPos+j][MEDIAN_RADIUS+i],
                    dataB[smemPos+j][MEDIAN_RADIUS+i],
                    dataR[smemPos+m][MEDIAN_RADIUS+n],
                    dataG[smemPos+m][MEDIAN_RADIUS+n],
                    dataB[smemPos+m][MEDIAN_RADIUS+n]
                );

        if(sumOfDistance < minDistance)
        {
            minDistance = sumOfDistance;
            clrX = dataR[smemPos+j][MEDIAN_RADIUS+i];
            clrY = dataG[smemPos+j][MEDIAN_RADIUS+i];
            clrZ = dataB[smemPos+j][MEDIAN_RADIUS+i];
        }
    }
```

Figure 5. The computation of the median vector is comprised of 4 nested loops.

2.4 MPI Implementation

To compare the performance results of the CUDA implementation, we chose to use MPI as the secondary development model. The program is executed with varying processor numbers. With the number of processors being N (1 master and N-1 slaves), the input image data is divided into N blocks and N-1 of these blocks are distributed among the accompanying slave processors (Fig. 6). After distribution is completed, the master and slave processors start computing the vector median filter. When slaves complete their assigned workloads, they send their output image data back to the master node. The outputs received from slaves are concatenated to obtain the final output image. In this parallel strategy, the slave processors do not need to communicate with each other, thus they only receive and send back a portion of the whole image.

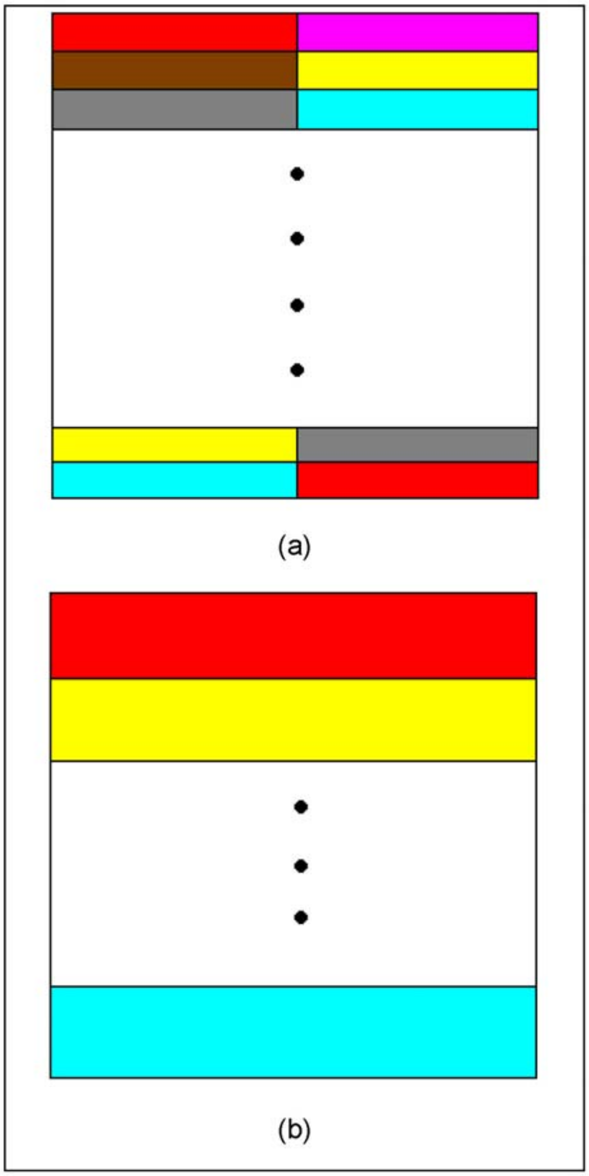


Figure 6. The comparison of parallel strategy of CUDA implementation vs. MPI implementation. (a) In the parallel strategy for CUDA, the input image is divided to equal sized tiles and assigned to individual thread blocks. (b) In MPI strategy, whole image is divided into relatively larger N parts that are distributed to accompanying processors.

3.0 DISCUSSION

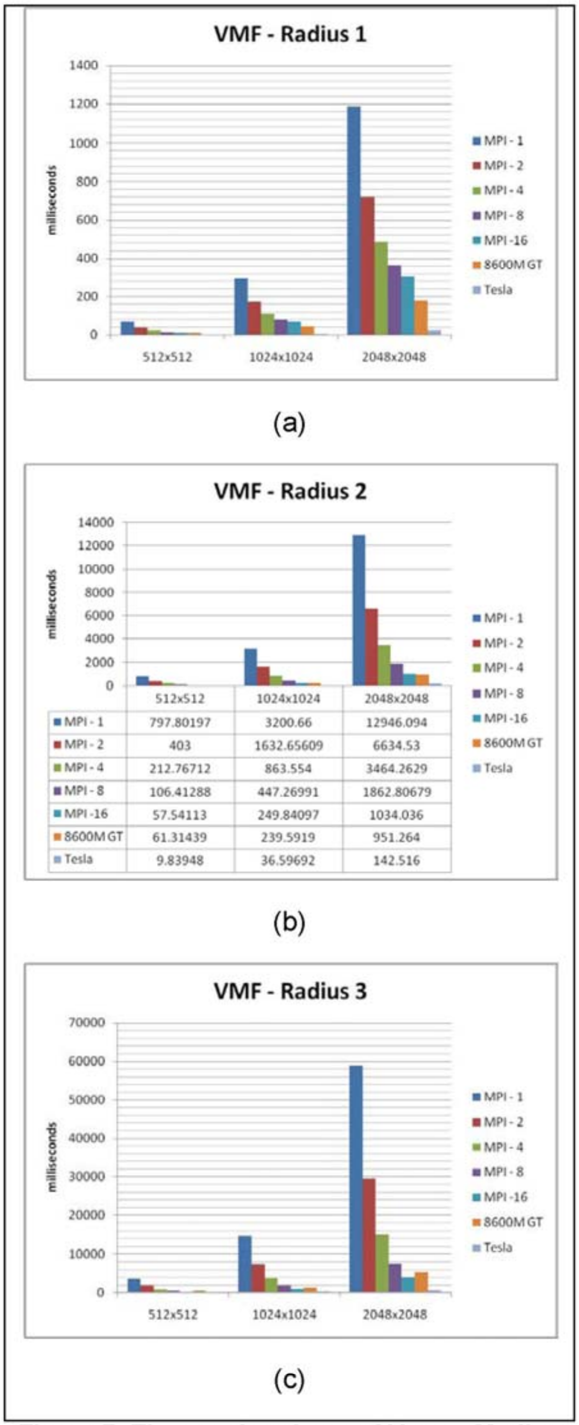
We chose different platforms to compare the performance of vector median filter. The two GPU platforms are the NVIDIA GeForce 8600M GT on a laptop computer and NVIDIA Tesla Personal Supercomputer on a workstation machine, whose specifications are given in Table 1.

Table 1. 8600M GT and TESLA GPU specifications.

	8600M GT	TESLA
# of SMs	4	30
# of Cores	32	240
Global Memory	256 MB	4 GB
Memory Bandwidth	9.25 GB/sec	102 GB/sec
Clock Rate	0.95 GHz	1.3 GHz

MPI parallel programming model is used as the platform of choice for sequential (single processor) and parallel CPU implementations. MPI code is run on Old Dominion University's High Performance Computing Cluster "Zorka" that has a number of Quad core 2.693 GHz AMD Opteron processors. In our experiments, we utilized 1, 2, 4, 8, and 16 processors.

The performance analysis of vector median filter is performed with varying image and kernel window radius sizes. Input images are 24-bit RGB images at the resolution of 512 x 512, 1024 x 1024, and 2048 x 2048 pixels. The applied kernel radius' are 1 (3 x 3), 2 (5 x 5), and 3 (7 x 7). Figure 7 presents the comparison results. Tesla platform outperformed MPI implementations in every case. We obtained close performance results when we utilized 16 MPI processors against the 8600M GT GPU. In other processor settings, the 8600M GT GPU outperformed MPI implementations.



(a)

	512x512	1024x1024	2048x2048
MPI - 1	797.80197	3200.66	12946.094
MPI - 2	403	1632.65609	6634.53
MPI - 4	212.76712	863.554	3464.2629
MPI - 8	106.41288	447.26991	1862.80679
MPI -16	57.54113	249.84097	1034.036
8600M GT	61.31439	239.5919	951.264
Tesla	9.83948	36.59692	142.516

(b)

(c)

Figure 7. The running times of Vector Median Filter (in milliseconds) on different platforms with varying image resolutions.

4.0 CONCLUSION

Vector median filter is a long used effective noise eliminating filter. In addition to removing noise and other artifacts, it is also capable of preserving edges and important features in an image. One drawback of vector median filter is its computational inefficiency. In this work, we demonstrated the extensive computing power of GPUs and harnessed this power to accelerate the Vector Median Filter computations with NVIDIA's Compute Unified Device Architecture (CUDA). For comparison purposes, filter was also implemented by using Message Passing Interface and run on a high performance computing cluster with varying processor numbers.

CUDA implementations were superior in most of the cases. The MPI implementation was faster than the 8600M GT laptop GPU when it was using 16 processors and working on smallest image size (512 x 512). The average speed-ups of Tesla GPU over MPI implementations are presented in Figure 8.

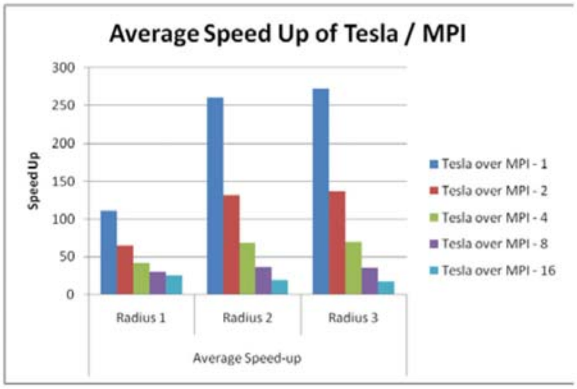


Figure 8. The average speed-ups of Tesla implementation over MPI implementations.

We have ported the computationally intensive Vector Median Filter to massively parallel GPU environment by using NVIDIA's CUDA. In order to obtain the maximum achievable performance, a CUDA developer not only has to take into account many parameters, but also has to rethink the algorithm in a parallel fashion. Although we believe that we produced decent results for the vector median filter using CUDA, a near-term goal is to further optimize this implementation.

5.0 REFERENCES

[1] T. Sun and Y. Neuvo, "Detail-preserving median based filters in image processing," *Pattern Recognition Letters,* vol. 15, pp. 341-347, 1994.

[2] V. Barnett, "The ordering of multivariate data," *Journal of the Royal Statistical Society. Series A,* vol. 139, pp. 318-355, 1976.

[3] I. Pitas and P. Tsakalides, "Multivariate ordering in color image filtering," *Circuits and Systems for Video Technology, IEEE Transactions on,* vol. 1, pp. 247-259, 295-6, 1991.

[4] J. Astola, P. Haavisto, and Y. Neuvo, "Vector median filters," *Proceedings of the IEEE,* vol. 78, pp. 678-689, 1990.

[5] M. Barni and V. Cappellini, "On the computational complexity of multivariate median filters," *Signal Processing,* vol. 71, pp. 45-54, 1998.

[6] R. J. Rost, *OpenGL(R) Shading Language (2nd Edition)*: Addison-Wesley Professional, 2005.

[7] Khronos Group, *The OpenCL Specification Version 1.0*. Khronos Group, 2009.

[8] NVIDIA, "NVIDIA CUDA C Programming Guide," 2010.

[9] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, Washington, DC, USA, 1998, p. 839.

[10] K. Oistamo, Q. Liu, M. Grundstrom, and Y. Neuvo, "Weighted vector median operation for filtering multispectral data," in *Systems Engineering, 1992., IEEE International Conference on*, 1992, pp. 16-19.

[11] Y. Shen and K. E. Barner, "Fuzzy vector median-based surface smoothing," *Visualization and Computer Graphics, IEEE Transactions on,* vol. 10, pp. 252-265, 2004.

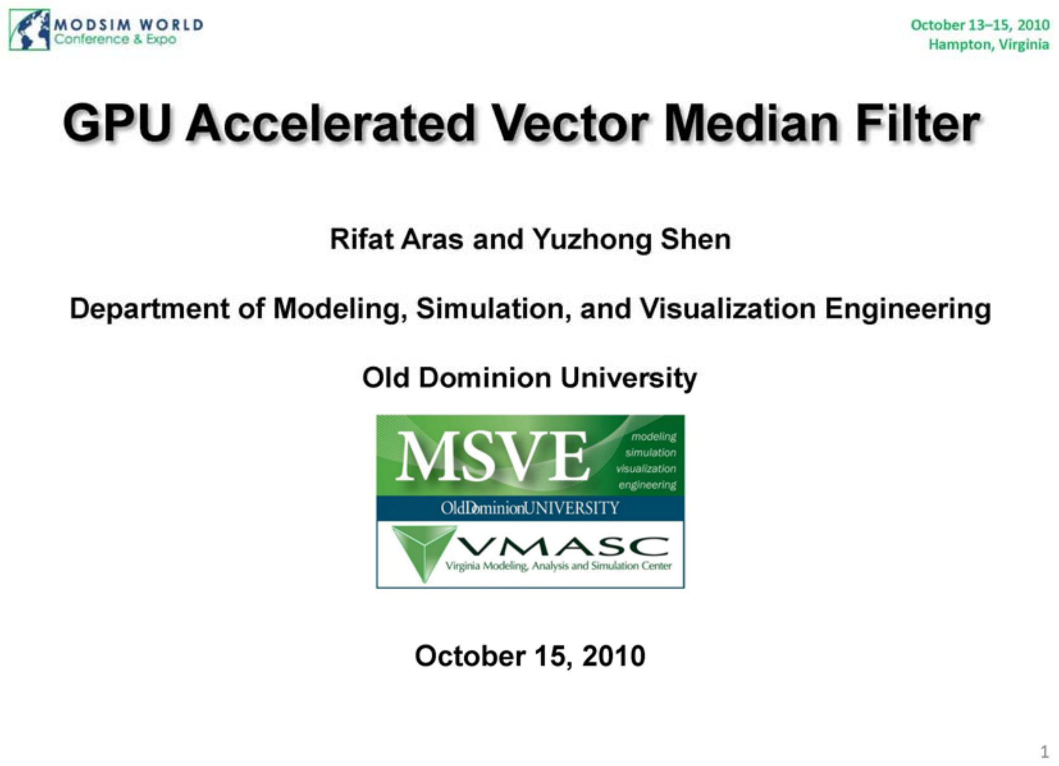
[12] A. Boudabous, L. Khriji, A. B. Atitallah, P. Kadionik, and N. Masmoudi, "Efficient architecture and implementation of vector median filter in co-design context," *Radio Engineering,* vol. 16, pp. 113-119, 2007.

[13] M. Barni and V. Cappellini, "A computationally efficient implementation of the L_1 vector median filter," in *Digital Signal Processing Proceedings, 1997. DSP 97., 1997 13th International Conference on*, 1997, pp. 283-286 vol.1.

[14] D. B. Kirk and W.-m. W. Hwu, *Programming Massively Parallel Processors*, 1 ed.: Elsevier, 2010.

[15] M. Pharr and R. Fernando, *Gpu Gems 2: Programming techniques for high-performance graphics and general-purpose computation*: Addison-Wesley Professional, 2005.

[16] K. Fatahalian and M. Houston, "GPUs: A closer look," *Queue,* vol. 6, pp. 18-28, 2008.



MODSIM WORLD Conference & Expo

October 13–15, 2010
Hampton, Virginia

GPU Accelerated Vector Median Filter

Rifat Aras and Yuzhong Shen

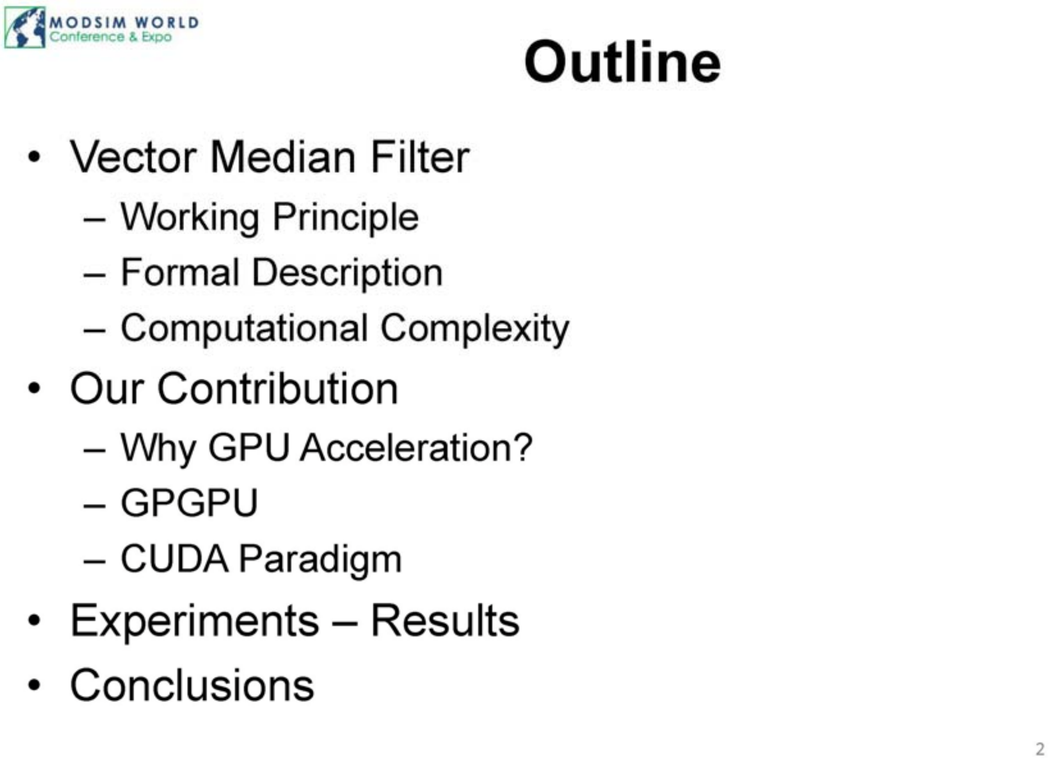
Department of Modeling, Simulation, and Visualization Engineering

Old Dominion University

MSVE modeling simulation visualization engineering
Old Dominion UNIVERSITY
VMASC Virginia Modeling, Analysis and Simulation Center

October 15, 2010

1



MODSIM WORLD Conference & Expo

Outline

- Vector Median Filter
  - Working Principle
  - Formal Description
  - Computational Complexity
- Our Contribution
  - Why GPU Acceleration?
  - GPGPU
  - CUDA Paradigm
- Experiments – Results
- Conclusions

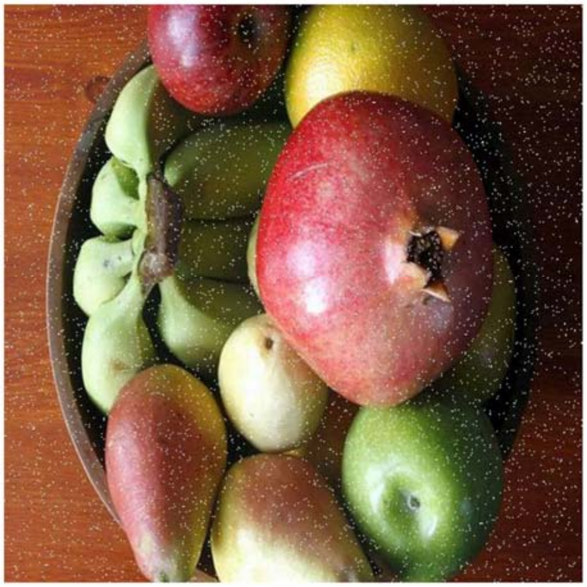
2

Vector Median Filter

- An effective noise elimination step for image processing tasks

T. Sun and Y. Neuvo, "Detail-preserving median based filters in image processing,“ *Pattern Recognition Letters,* vol. 15, pp. 341-347, 1994.

3

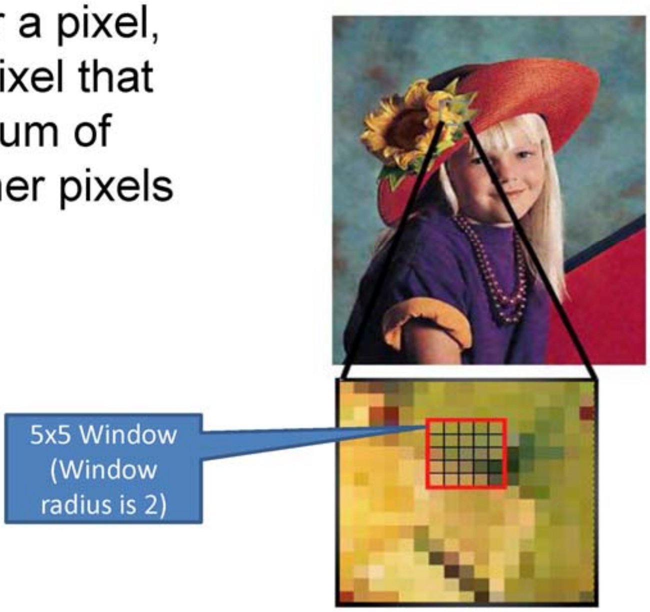


4

Preserving edges, lines, and other image structures while removing noise artifact

Working Principle

- For a window over a pixel, VMF selects the pixel that has the smallest sum of distance to the other pixels in the window.



5x5 Window (Window radius is 2)

Formal Description

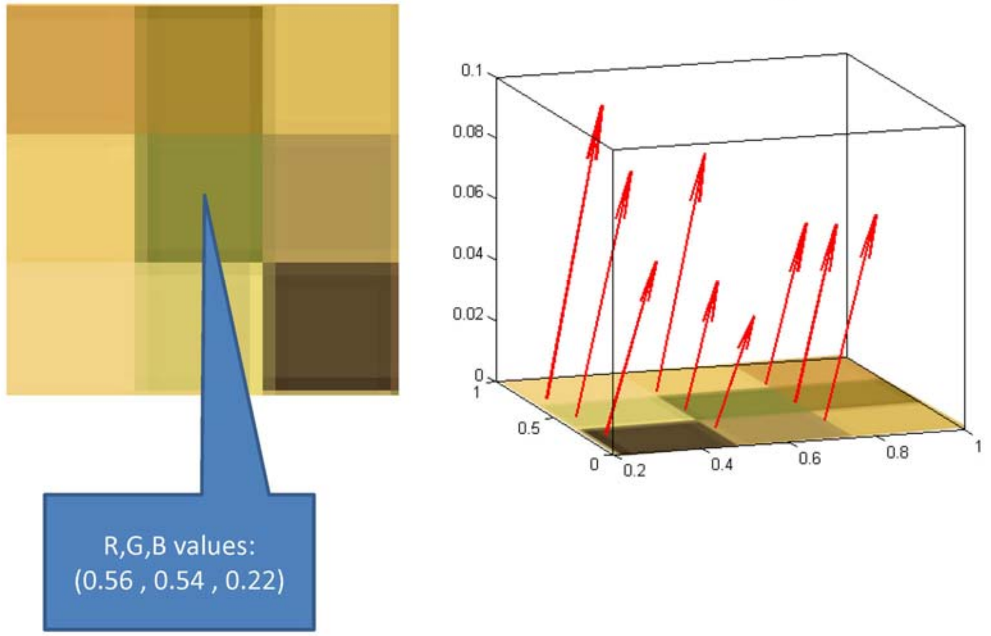
Given a window that contains N = nxn pixels denoted by W={x_1,x_2,…,x_N} the output of vector median filter x_{VM} is computed by

$$\sum_{i=1}^{N} \|x_{VM} - x_i\| \le \sum_{i=1}^{N} \|x_j - x_i\|, j = 1, \dots, N, (1)$$

||.|| denotes the distance metric between two vectors. In this work, Euclidean distance (L2-norm) is used.

$$\|u - v\| = \sqrt{\sum_{k=1}^{p} (u_{(k)} - v_{(k)})^2}. \qquad (2)$$

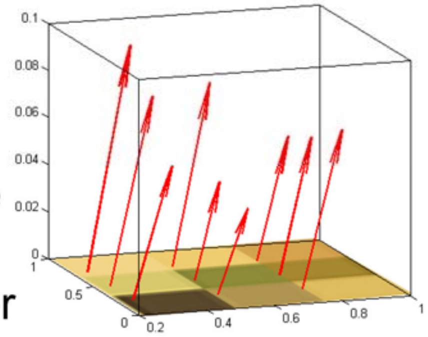
Vectors in an Image



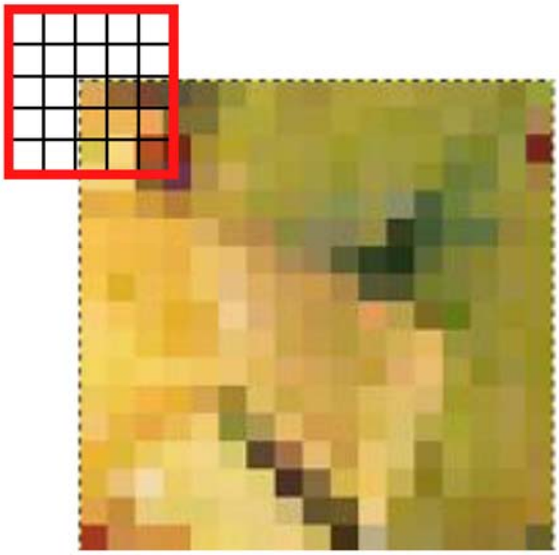
R,G,B values:
(0.56 , 0.54 , 0.22)

Computational Complexity

- For the nxn window case, each n^2 vector is compared to other n^2-1 vectors.
- Each Euclidean distance calculation involves 8 floating point operations and a square root.
- For 3x3 case, the total number of operations is equal to 576 floating points and 72 square roots.



Sequential Implementation



MODSIM WORLD Conference & Expo

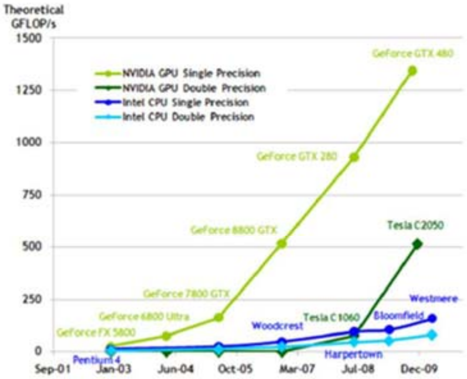
Main Contribution

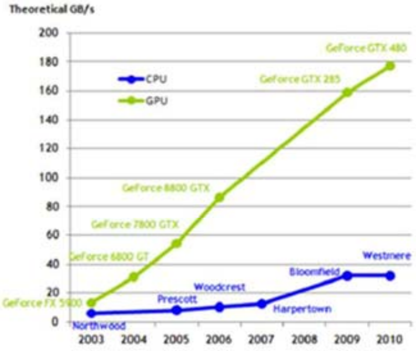
- Implementing VMF using the CUDA programming paradigm
- Comparing the performance of GPU accelerated implementation to the single thread CPU and multi-processor MPI versions

11

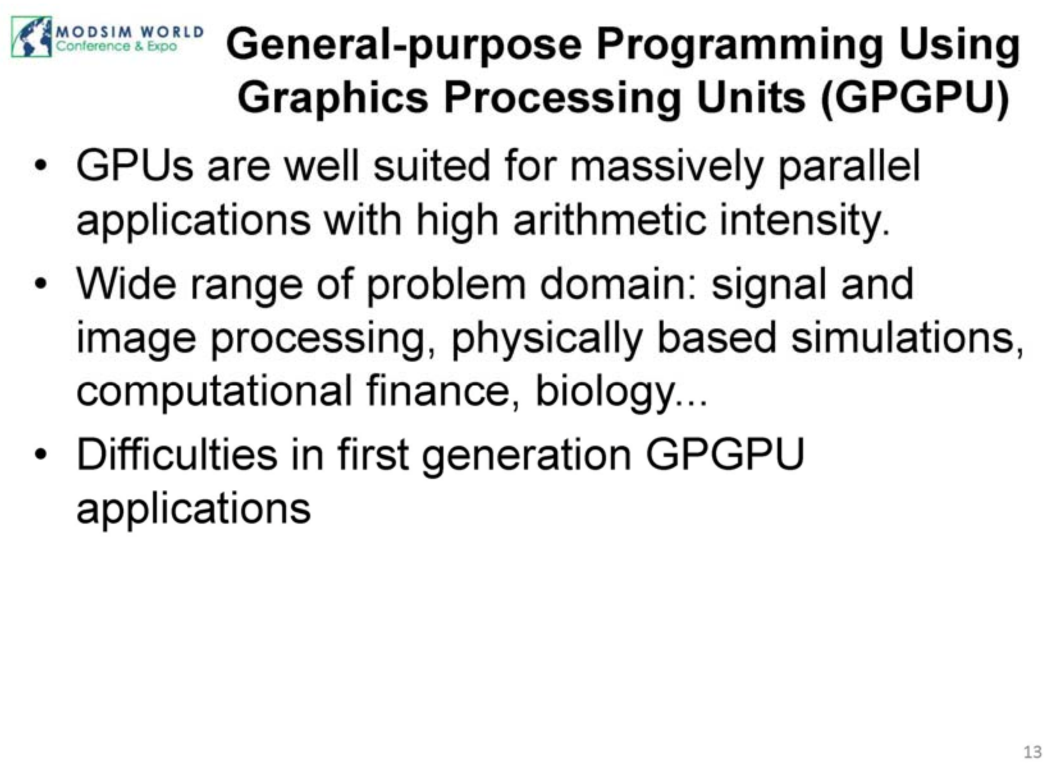
MODSIM WORLD Conference & Expo

Why GPU Acceleration?





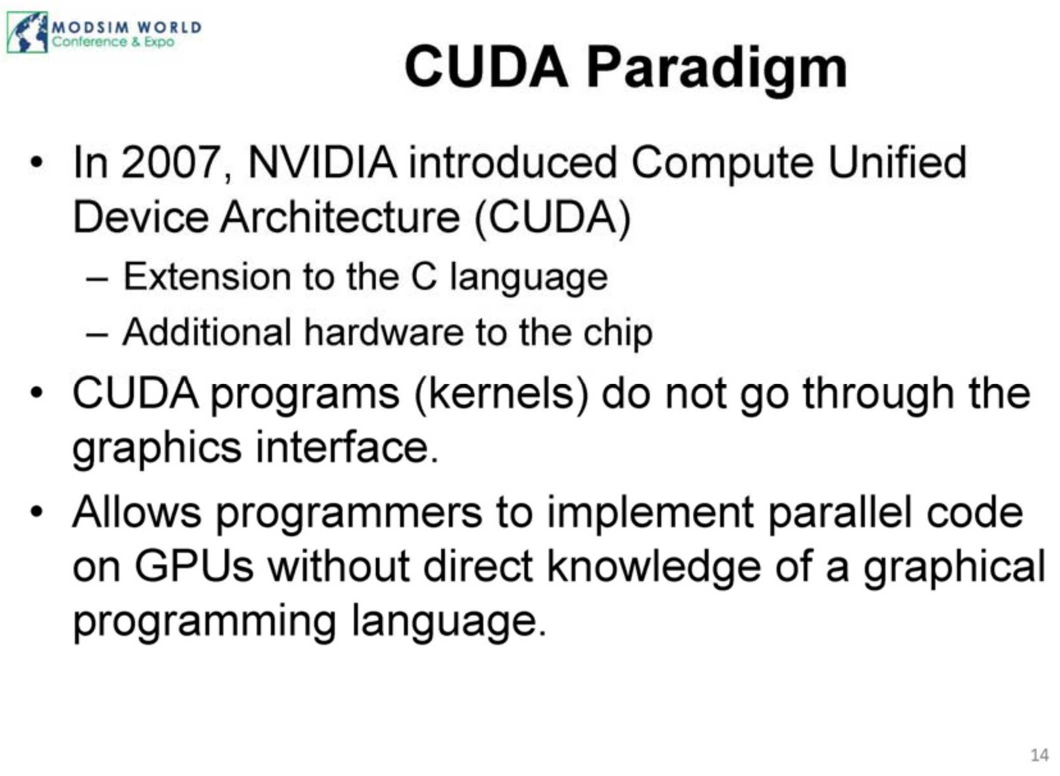
12



General-purpose Programming Using Graphics Processing Units (GPGPU)

- GPUs are well suited for massively parallel applications with high arithmetic intensity.
- Wide range of problem domain: signal and image processing, physically based simulations, computational finance, biology...
- Difficulties in first generation GPGPU applications

13



CUDA Paradigm

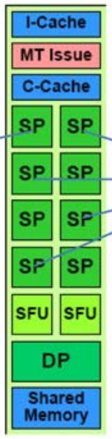
- In 2007, NVIDIA introduced Compute Unified Device Architecture (CUDA)
  - Extension to the C language
  - Additional hardware to the chip
- CUDA programs (kernels) do not go through the graphics interface.
- Allows programmers to implement parallel code on GPUs without direct knowledge of a graphical programming language.

14

CUDA Enabled GPU Architectures

- Arrays of Streaming Multiprocessors (SMs)

Each core has:
- A multiply-add unit
- An additional multiply unit
- Special function FLOP units



8 streaming processors (cores)

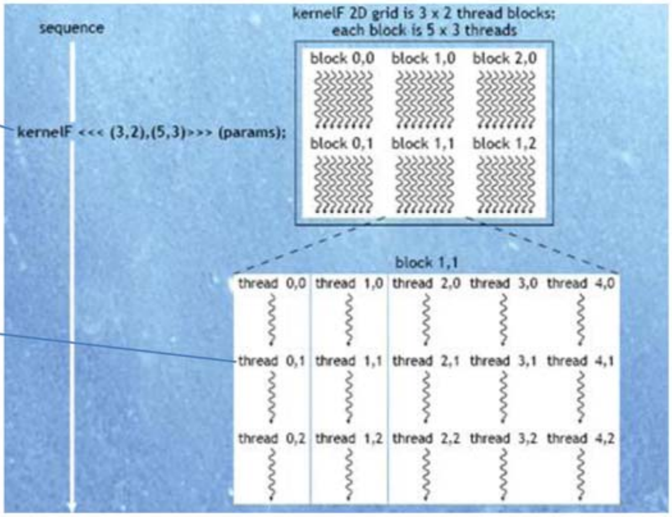
Cores of the same SM share:
- Control logic
- Instruction cache
- Fast access shared memory

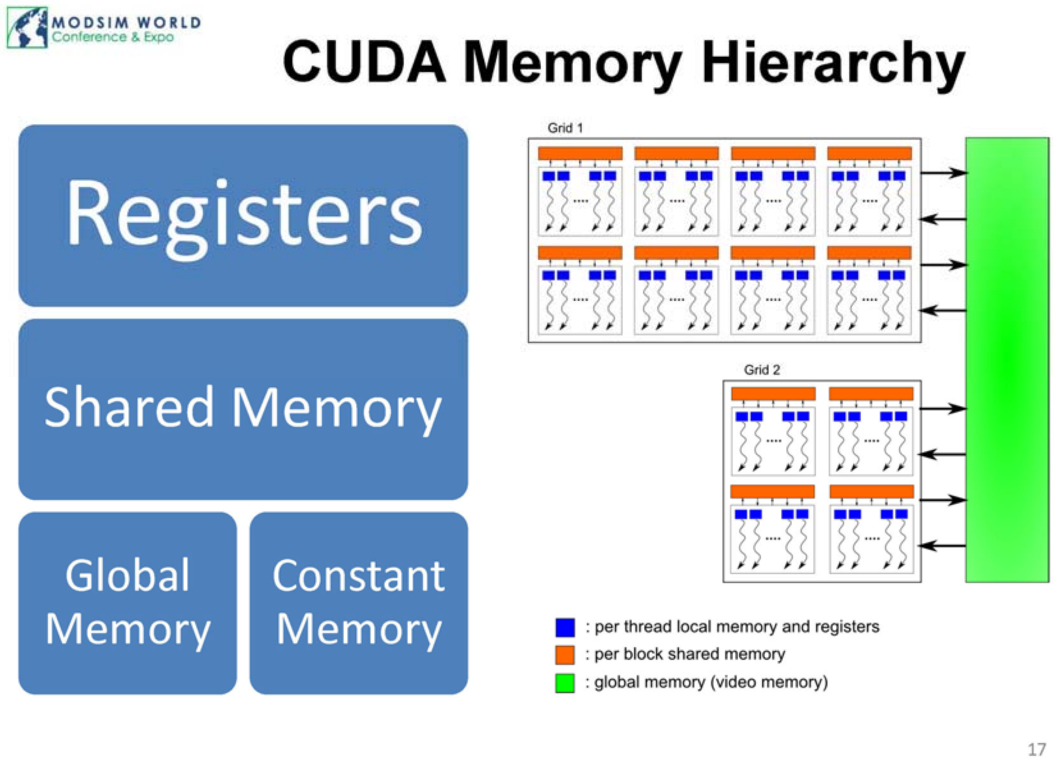
GPUs with GT200 architecture can support 1024 concurrent threads per SM.

Fine-grained Data Parallelism

Kernels: User created functions that are executed by each individual thread

Threads are organized into a two-level hierarchy:
- 2D grid consisting of blocks
- 3D blocks consisting of threads





MODSIM WORLD Conference & Expo

CUDA Memory Hierarchy

Registers

Shared Memory

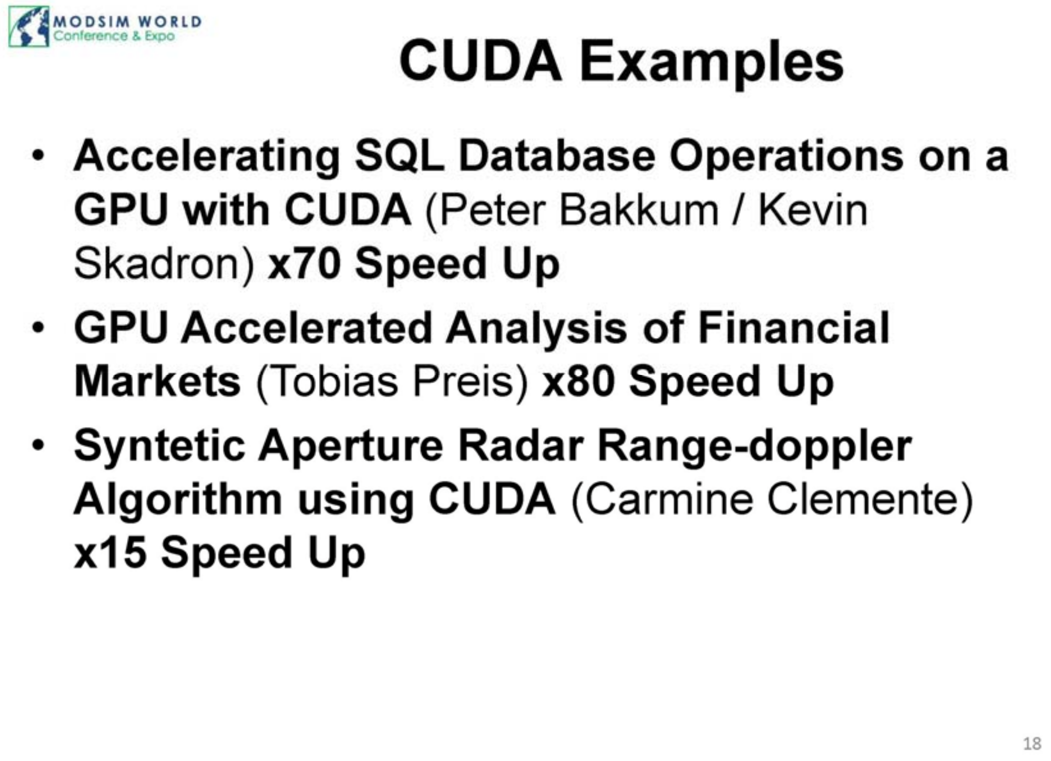
Global Memory

Constant Memory

🟦 : per thread local memory and registers

🟧 : per block shared memory

🟩 : global memory (video memory)



MODSIM WORLD Conference & Expo

CUDA Examples

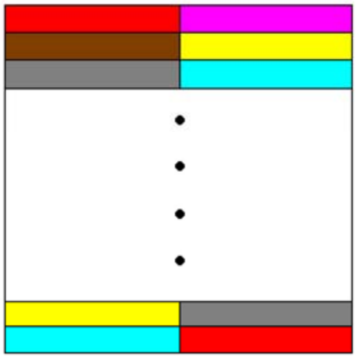
- **Accelerating SQL Database Operations on a GPU with CUDA** (Peter Bakkum / Kevin Skadron) **x70 Speed Up**
- **GPU Accelerated Analysis of Financial Markets** (Tobias Preis) **x80 Speed Up**
- **Syntetic Aperture Radar Range-doppler Algorithm using CUDA** (Carmine Clemente) **x15 Speed Up**

Experiment Setup

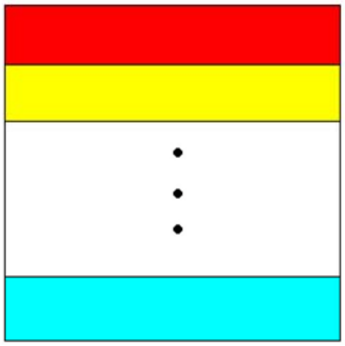
- 3 different image sizes: 512x512, 1024x1024, and 2048x2048
- 3 different window radius: 1(3x3), 2(5x5), and 3(7x7)
- Two GPU platforms
  - NVIDIA 8600M GT (Mobile GPU)
  - NVIDIA Tesla Personal Supercomputer
- 1, 2, 4, 8, and 16 MPI Processors (Quad core 2.693 GHz AMD Opterons).

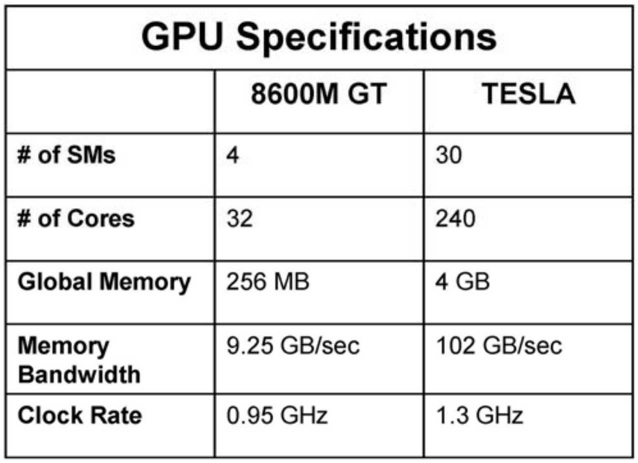
Parallel Strategies

CUDA Strategy



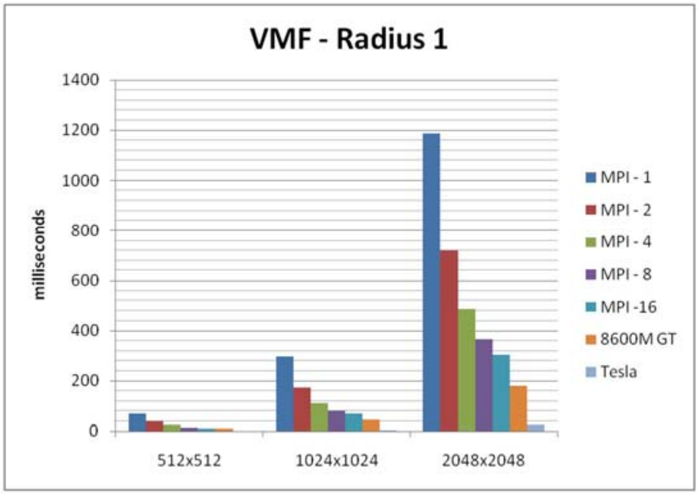
MPI Strategy



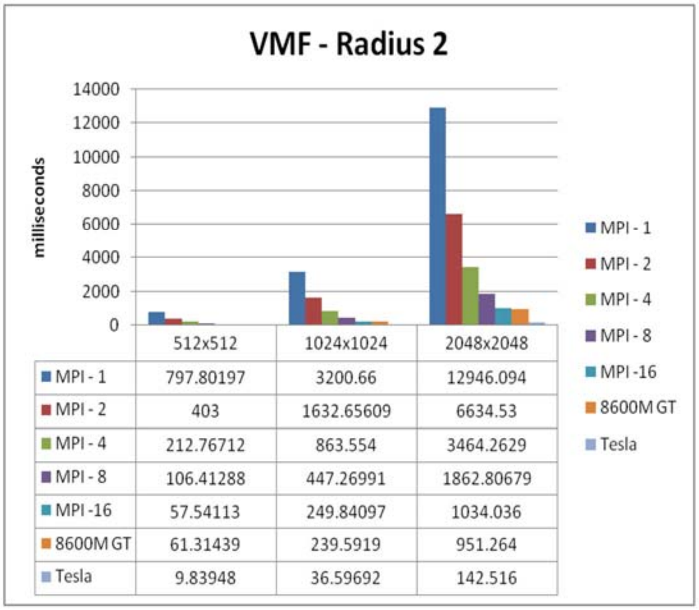


GPU Specifications		
	8600M GT	**TESLA**
# of SMs	4	30
# of Cores	32	240
Global Memory	256 MB	4 GB
Memory Bandwidth	9.25 GB/sec	102 GB/sec
Clock Rate	0.95 GHz	1.3 GHz

Results – VMF Window Size: 3x3



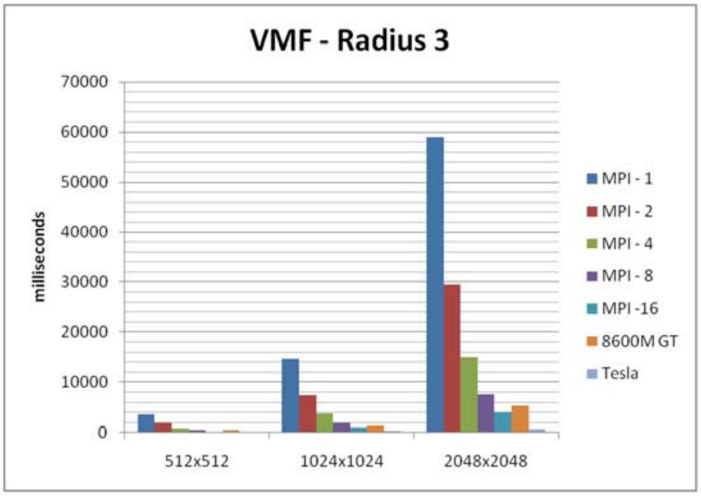
Results – VMF Window Size: 5x5



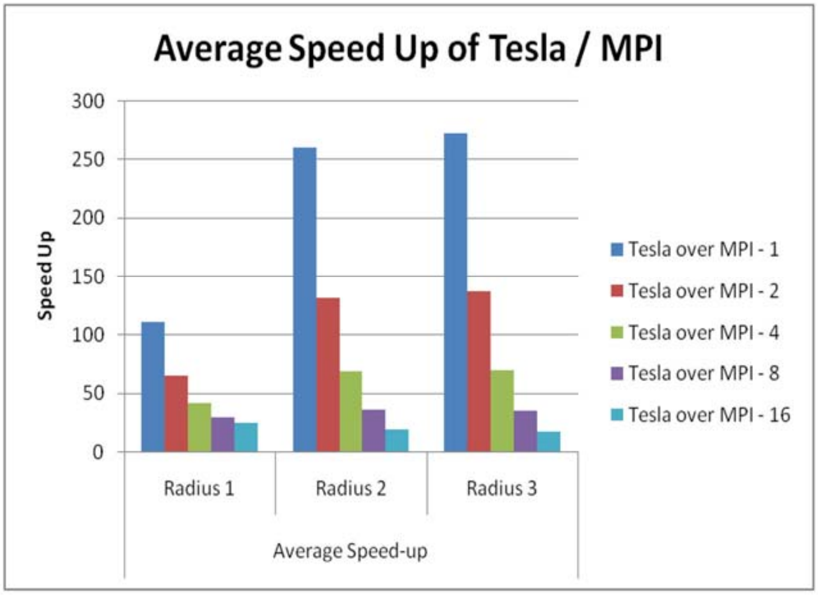
	512x512	1024x1024	2048x2048
MPI - 1	797.80197	3200.66	12946.094
MPI - 2	403	1632.65609	6634.53
MPI - 4	212.76712	863.554	3464.2629
MPI - 8	106.41288	447.26991	1862.80679
MPI -16	57.54113	249.84097	1034.036
8600M GT	61.31439	239.5919	951.264
Tesla	9.83948	36.59692	142.516

23

Results – VMF Window Size: 7x7

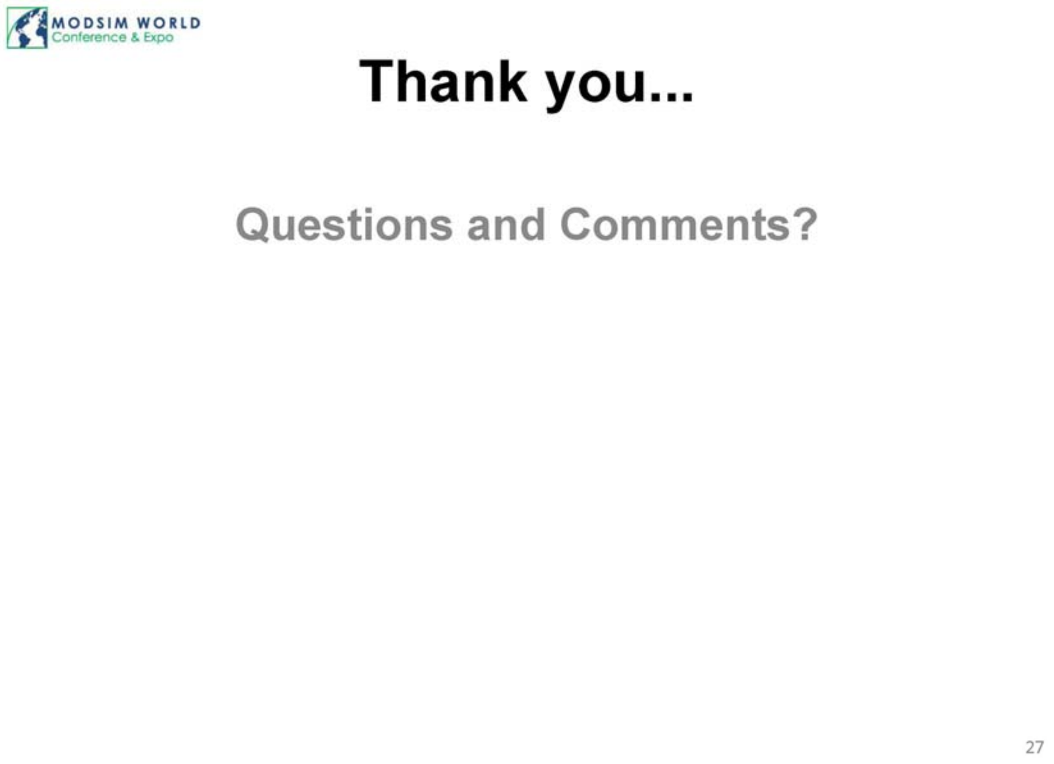


24



Conclusions

- Vector Median Filter: Removing noise, preserving edges and other important image features
- High computational complexity
- We have ported VMF to massively parallel GPU environment using CUDA.
- Obtained 100x – 275x of speed-up over single processor implementation



MODSIM WORLD
Conference & Expo

Thank you...

Questions and Comments?

27

2.5 Agent-Based Simulations for Project Management

Agent-Based Simulations for Project Management

J. Chris White
ViaSim Solutions
Rockwall, TX
jcwhite@viasimsolutions.com

Robert M. Sholtes
SimBLOX, LLC
Rockville, MD
rsholtes@simblox.com

Abstract: Currently, the most common approach used in project planning tools is the Critical Path Method (CPM). While this method was a great improvement over the basic Gantt chart technique being used at the time, it now suffers from three primary flaws: (1) task duration is an input, (2) productivity impacts are not considered, and (3) management corrective actions are not included. Today, computers have exceptional computational power to handle complex simulations of task execution and project management activities (e.g., dynamically changing the number of resources assigned to a task when it is behind schedule). Through research under a Department of Defense contract, the author and the ViaSim team have developed a project simulation tool that enables more realistic cost and schedule estimates by using a resource-based model that literally turns the current duration-based CPM approach "on its head." The approach represents a fundamental paradigm shift in estimating projects, managing schedules, and reducing risk through innovative predictive techniques.

1.0 BACKGROUND

For both commercial and government organizations, the ability to manage projects effectively is a major contributor to an organization's overall performance. If an organization cannot manage its internal projects effectively, resources, time, and money are wasted. For commercial organizations, this weakens the organization's market position and capacity to generate business, which can ultimately result in closing down entire business units within the organization or the entire organization. For government and military organizations, this reduces the effectiveness of the organization's operations and system development efforts, which can ultimately result in the loss of human lives. Therefore, effective project management (PM) is necessary for strong organizational performance. This is not news to anyone.

However, times are changing for the project manager. Welcome to the 21st century! Projects are more complex than ever, tough economic conditions put enormous pressure on achieving success, and the typical project manager has several concurrent projects and no longer has the luxury of focusing on only one project. In addition, today's computers are capable of handling more computationally intensive analyses, so why not use this power with an agent-based simulation approach to project management?

This paper will review the current approaches used in today's project management tools and introduce a new, agent-based simulation approach that offers a higher level of realism and, consequently, a higher level of project success.

2.0 CURRENT PROJECT MANAGEMENT (PM) METHODS

Almost all project planning and scheduling tools on the market today use some type of PERT and/or CPM methodology as their primary underlying methodology. PERT (Program Evaluation and Review Technique) was invented by the U.S. Navy in the 1950's to manage the Polaris submarine missile program. CPM (Critical Path Method) was invented about the same time in the private sector. These two approaches are synonymous and are often interchanged or even collectively called PERT/CPM.

The PERT/CPM approach (Figure 1) is considered superior to the previously preferred Gantt approach, a horizontal bar chart developed as a production control tool in 1917 by Henry L. Gantt. A major problem with the Gantt method was that it did not indicate task dependencies so a PM could not tell how one task falling behind schedule affects other tasks in the project. The PERT/CPM method is designed to do this. (It should be noted that today's Gantt charts actually leverage some of the PERT/CPM information by now showing dependency connections on the horizontal bar charts.)

PERT charts depict task, duration, and dependency information. Each chart starts with an initiation node from which the first task (or tasks) originates. If multiple tasks begin at the same time, they are all started from the node or branch, or fork out from the starting point. Each task is represented by a line which states its name or other identifier, its duration, the number of people assigned to it, and in some cases the initials of the personnel assigned. The other end of the task line is terminated by another node which identifies the start of another task, or the beginning of any slack time or float time (i.e., waiting time between tasks).

Each task is connected to its successor tasks in this manner forming a network of nodes and connecting lines. The chart is complete when all final tasks come together at the completion node. The key difference with the CPM method is that the CPM method highlights the critical path, the sequence of activities for which there is no (or the least amount of) slack or float time. Thus, by definition, the critical path is the pathway of tasks on the network diagram that has no extra time available (or very little extra time). Note that it is possible to have multiple critical paths.

The latest innovation in the PM world is the Critical Chain method developed by Eliyahu M. Goldratt in 1997. The critical "chain" is akin to the critical "path" of CPM, but has a slightly different definition: the path of dependent tasks that define the expected lower limit of a project's possible completion time. The Critical Chain method leans heavily on Parkinson's Law, which suggests that work will expand to fit the time allowed for it. It is believed that if you "hide" the extra time, people on the project won't know it's there and will work to meet the shorter task times. "Buffer" activities are then inserted into the project plan as buckets of extra time (i.e., collections of slack or extra time from the other activities). The purpose of the buffers is to protect the promised completion date from variation in the critical chain. In essence, the PM makes participants think it's a very short project, all the while keeping some extra time set aside in case any task runs late. It's more a psychological game than a true management method. Thus, it is not much of an innovation. The underlying approach, PERT/CPM, is still the same fundamental approach. To be sure, the Critical Chain method helps a little, but the method is still inhibited by insufficiencies inherent within the underlying PERT/CPM methodology.

3.0 LIMITATIONS OF CURRENT PM METHODOLOGIES

To date, improvements to PM tools have been evolutionary and not revolutionary. All the PM tools are still based on the PERT/CPM approach. The improvements that occur with each round of new PM tools are merely "bells and whistles" (e.g., better tracking of resources, uploading from spreadsheets) or perhaps enabling the tool to be used over the internet. Unfortunately, a 50-year old methodology delivered over the internet is still a 50-year old methodology.

Fundamentally, the PERT/CPM approach suffers from three (3) major flaws:

1. Task duration is an input. However, many factors (e.g., availability and productivity of resources, dependencies among tasks, hours worked by employees) affect the duration of a task. Thus, in the real world, task duration is actually an output.
2. Productivity impacts are not considered. In current PM tools, labor can

be added to or removed from a task with no impact on the productivity of labor applied to the task. Today's tools assume all resources are equal. Yet, we know they are not. New employees or junior-level employees do not get as much work done in the same period of time as experienced, senior-level employees. Also, in current PM tools people can be scheduled for overtime with no impact on their productivity. However, anyone who has worked a significant amount of overtime can validate that productivity decreases due to fatigue or burnout. Working a little overtime on a couple of days usually has a negligible impact, but long durations of working overtime can have significant impacts on labor productivity. Lastly, it is a known fact (especially on software development projects) that throwing more people at a task often makes the task fall further behind schedule due to lower labor productivity as experienced people train the new people and the new people make mistakes that must be corrected.

3. Corrective actions are not captured. The actual management decisions and actions that PM's take during a project are not included. However, these corrective actions can significantly influence progress. Current tools only match resources against task assignments. As a result, current PM tools allow for static planning, but not dynamic reaction and re-planning. In current PM tools, if it looks like a task will run late (e.g., based on the Earned Value schedule performance index, SPI), the project manager must develop several different plans through trial-and-error to see if they will work. The current tools do not help the project manager actually manage the project. The tools only allow the PM to develop multiple, static plans with no insight.

These inherent flaws indicate that our current PM tool set is too simplistic and does not reflect reality. As a result, these tools cause us to make decisions that are detrimental to project success. In fact, it is not uncommon to doom ourselves to failure (or at least a very long and difficult road) with our first baseline project plan. In other words, right out of the gate we are already off course! This is not a criticism of project managers; it is a criticism of the simplistic approaches found in today's PM tools that give us insufficient and sometimes even incorrect answers. Consequently, we often rely heavily on individual PM's to single-handedly make projects successful. We applaud heroic efforts where PM's work around the "system" to make everything work out right. Why not have a "system" that actually helps the PM succeed?

Essentially, the current PM tools are not capable of handling the complexity of the issues experienced on most projects because they are rooted in a simplistic approach that was developed 50 years ago. When it was developed, the PERT/CPM approach was innovative and useful. Unfortunately, its effectiveness and appropriateness have significantly eroded over time. If we want a better tool for PM's, we need a better approach than PERT/CPM.

4.0 OVERVIEW OF THE DYNAMIC PROGRESS METHOD (DPM)

The Dynamic Progress Method (DPM) is a new approach to planning, estimating, and managing projects that builds upon the power now found in computers and applies a different type of simulation model than is currently used in most tools. The underlying simulation model in most of today's project management (PM) tools is very simplistic because it was developed at a time when computers had very little computational power. Thus, the model had to be simple enough to use a pencil-and-paper approach as a back-up. Now, computers have much greater power…so let's use it. Furthermore, we've learned a lot more about managing projects, so let's put that knowledge to use, too.

The goal of DPM is to address the three major flaws of the current CPM approach. DPM begins with a fundamental re-evaluation of how input information is used. Figure 2 provides a good example of this. After defining which tasks are to be done and their dependencies, a project manager (PM) then begins the process of estimating for each task. For a task, the PM usually knows about how much work needs to be done for the task, who might do that work, an idea as to how "good" or "effective" that person(s) is, and an idea about the availability of that person. In the example in Figure 2, it's easy to see that an 80-hour task with one assigned resource that is available 8 hours/day and is 100% productive (i.e., for each hour the resource is paid, the resource completes an hour of actual project work) gives a task duration of 10 days [(80 hours) / (8 hours/day) = 10 days]. With CPM-based planning tools, the duration is the input. Conversely, with DPM-based tools, the individual task and resource inputs are used instead. In this example, if the resources are actually available at the level of productivity assumed, then the duration will be the same 10 days.

In the previous section, we discussed how having task duration as in input is a flaw in today's tools. Let's use this example to highlight this fact. In this example, notice that the duration input for CPM does not need any of the detailed task/resource information. As a result, a project manager can input a duration without knowing the details. Any duration is just as good as any other duration. Some people may think that this is a good thing because it simplifies the data input process by not burdening the PM with figuring out all those details. However, that fact that an entire plan can be constructed without the resource details is a dangerous endeavor. How can these estimates be justified and verified? They can't! Yet, a great deal of trust is place on these plans that are tenuous at best. DPM requires the PM to be explicit about the estimates. If a duration is assumed, why? Surely there must be some information that the PM is basing the estimate on? DPM pushes the project manager to use that information. It's another level of detail. It seems at times that we are not willing to add this level of detail at the beginning of a project, yet during a project we are completely fine with re-working and re-planning a project multiple times at great expense (both time and money) as we realize that our estimates were insufficient (because they lacked detail).

What is interesting to note is that today's tools back-calculate some of these details. For example, if you input a duration of 10 days for a task and then assign a resource that has a 8-hour work day, the tool will back-calculate that the task is an 80-hour task. To test this, try adding another resource. With 2 people, the tool will automatically cut the duration down to 5 days [(80 hours) / (16 hours/day) = 5 days]. That is because the completion rate is doubled from 8 hours/day (1 person * 8 hours/day) to 16 hours/day (2 people * 8 hours/day). Thus, this information is actually used by the current tools, though it is not evident to the user. Or, change the work day for the single resource to 16 hours/day. You will get the same 5 day duration.

The underlying model for DPM is an operational model, which means that it mimics the actual actions and processes of a project. Figure 3 provides a schematic diagram of the task execution portion of the DPM model. (Note that everything in Figure 3 is actually contained in today's PM tools, just perhaps used a little differently.) DPM begins with a "bucket" of work that will be done for a task (Work To Do). As resources are allocated to the task, work is done (Completion Rate) and work moves from "to do" to "complete" (Completed Work). The rate at which work is completed is based on how many resources are allocated (Number of Resources) and how many hours per day those resources work (Actual Labor Hours). Then,

"Effective" Labor Hours = (Number of Resources) * (Actual Labor Hours)

Ex: "Effective" Labor Hours = (1 person) * (8 hours/person) = 8 hours

By changing the number or resources allocated to a task and/or the number of hours the resource is available to work on the task, the rate of work completion can be changed.

The graphics in the lower right of Figure 3 indicate that completion of work for a given task may also depend on the progression of work on other preceding tasks. For instance, Task 2 may require that Task 1 is 100% complete before Task 2 can start (i.e., Finish-to-Start dependency).

The term EV/Status Calculations at the top of Figure 3 indicates that at any point in time the status of the task relative to its expected schedule and cost can be determined. This shows whether a task is ahead/behind on schedule and over/under on cost. A growing trend is to use Earned Value metrics (e.g., Schedule Performance Index - SPI, Cost Performance Index - CPI) to represent "schedule pressure" and "cost pressure" on the task. (For a full discussion of Earned Value metrics, please see other sources.)

Where DPM adds value is shown in Figure 4 in green and red. One of the first new elements that DPM incorporates is Resource Productivity (i.e., the effectiveness or efficiency of a particular resource to get work done). Productivity is a measure of how much actual project work gets done for every hour of time the resource is paid. For example, as stated previously, new employees or junior-level employees do not get as much work done in the same period of time as experienced, senior-level employees. As another example, sometimes there is a "learning curve" associated with a project. A person may start the project with a low level of productivity because it is a new team and new work, but over time the person gains proficiency and grows to a higher level of productivity. DPM adds this productivity component.

Next, in the real world, status information (e.g., EV/Status Calculations) may drive a project manager to apply some sort of corrective action or management decision to get a task back on course. In the specific DPM model described here, there are two actions that a PM can take: (1) Add/remove resources, or (2) Add/remove work hours. For example, when a task is falling behind schedule, a PM may decide to allocate additional resources to the task, or the PM may decide to work everyone overtime, or the PM may decide to do a combination of both. However, in the real world there are consequences for these actions. There is not always a 1-for-1 return. For instance, as stated previously, anyone who has worked a significant amount of overtime can validate that productivity decreases due to fatigue or burnout. DPM incorporates this impact and it is shown by the red line with Fatigue.

Similarly, for a task there may be an "optimum" number of people to work on that task to maximize productivity. Having too few or too many people may make the resources less productive. In some cases, there may actually be physical limits that prevent throwing a lot of people at the job (e.g., limited space in a crawl space for electricians). Or, even though there may not be an optimum number of people to maximize productivity, it is common to experience productivity losses when more and more people are allocated to a job because of difficulties with communication and coordination among all the people. DPM incorporates this impact, also, and it is shown by the red line with Over-manning.

5.0 BENEFITS OF USING DPM

DPM has several major benefits over the current CPM approach.

- Proactive Approach: DPM allows PM's and others to review and challenge assumptions and plans before problems arise, which increases the probability of project success.

- Better Baseline Plans: DPM can help PM's launch their projects with more defendable and more achievable baseline plans.
- Better Corrective Actions: DPM can guide PM's on which corrective actions are most effective under which set of conditions. For instance, on one task it may be better to work resources overtime to accelerate the project, but on another task it may be better to add more resources.
- Real-World Consequences: Everyone knows that productivity can suffer when too many people are thrown at a job or people are working a lot of overtime. DPM incorporates these types of impacts.
- Better Risk Management: DPM can highlight the risks inherent in baseline or midstream estimates from changes at the resource level. This helps PM's know where to focus their attention.
- Project Acceleration: DPM can show realistic options for "accelerating" a project, along with cost and labor trade-offs for achieving the shorter schedule. DPM can even quantify the shortest possible duration for a project based on its current structure.

6.0 CASE STUDY: USING AN AGENT-BASED DPM SIMULATION TOOL

A project manager is leading a team to develop a prototype sensor system and is using Microsoft Project, which is a CPM-based tool. The baseline project plan (without leveling resources) shows a duration of 39 months at a total cost of $13.1M (Figure 5 and 6). To remove resource over-allocations, the project plan was level-loaded. After resource leveling week-to-week, Microsoft Project shows a duration of 599 months! Clearly, this is not realistic. When the resources are leveled month-to-month, Microsoft Project shows a duration of 77 months. It is difficult to trust these results when there is such a huge variance from minor changes.

The same three plans were imported from Microsoft Project into an agent-based DPM simulation tool with the assumption of 100% productivity for all resources. The results of all three simulations were identical: total duration 71 months with a total cost $13.9M. The DPM approach appears to provide a more consistent estimate. Standard PM tools tend to be either overly optimistic (when resources are not level-loaded) or overly pessimistic (when resources are level-loaded) using the simple CPM approach, whereas the DPM approach is realistic.

With the DPM approach, the PM also has the choice of changing the productivity of assigned resources. An additional simulation was run using a resource productivity level of 85%. This is a commonly assumed productivity level for labor resources. The simulation results from the agent-based DPM simulation tool for this scenario are a duration of 84 months and a total cost of $15.1M. This is the most realistic expectation for this project. The ability to bring simulation and productivity into the estimating process through the DPM approach makes initial project budgets and timelines more realistic than those developed in other CPM-based tools.

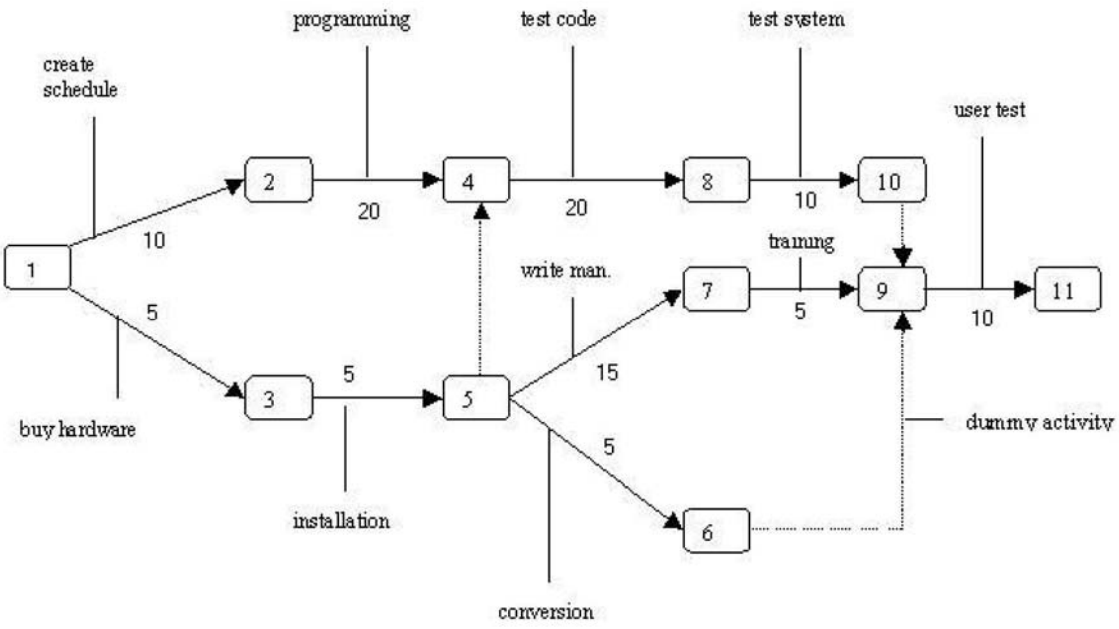
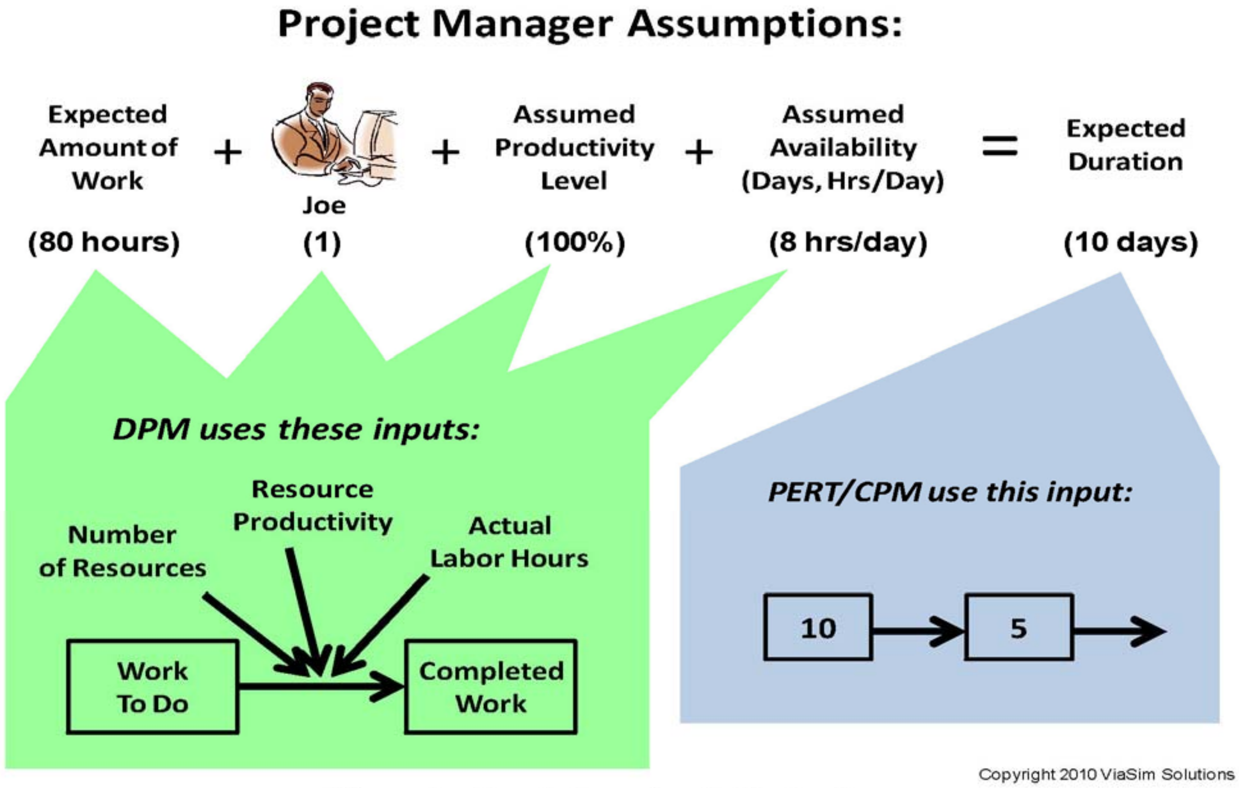


Figure 1: Simple PERT Diagram



Copyright 2010 ViaSim Solutions

Figure 2: How Information Is Used as Inputs

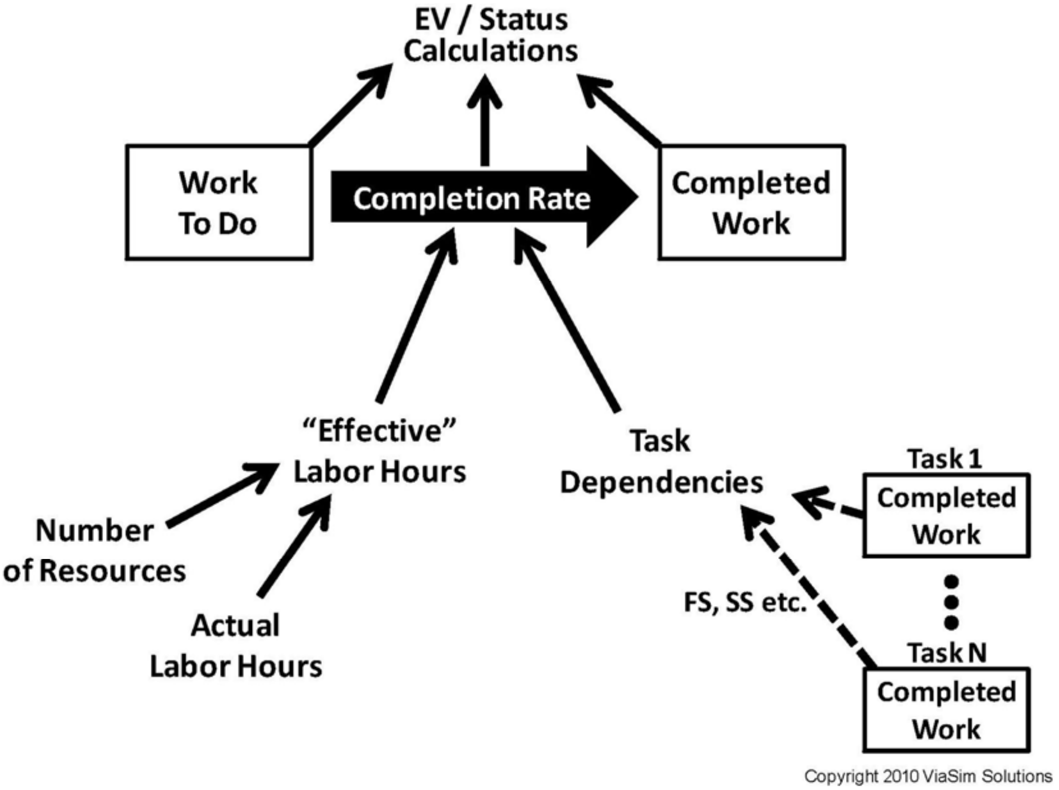


Figure 3: Overview of Operational Model for Task Execution

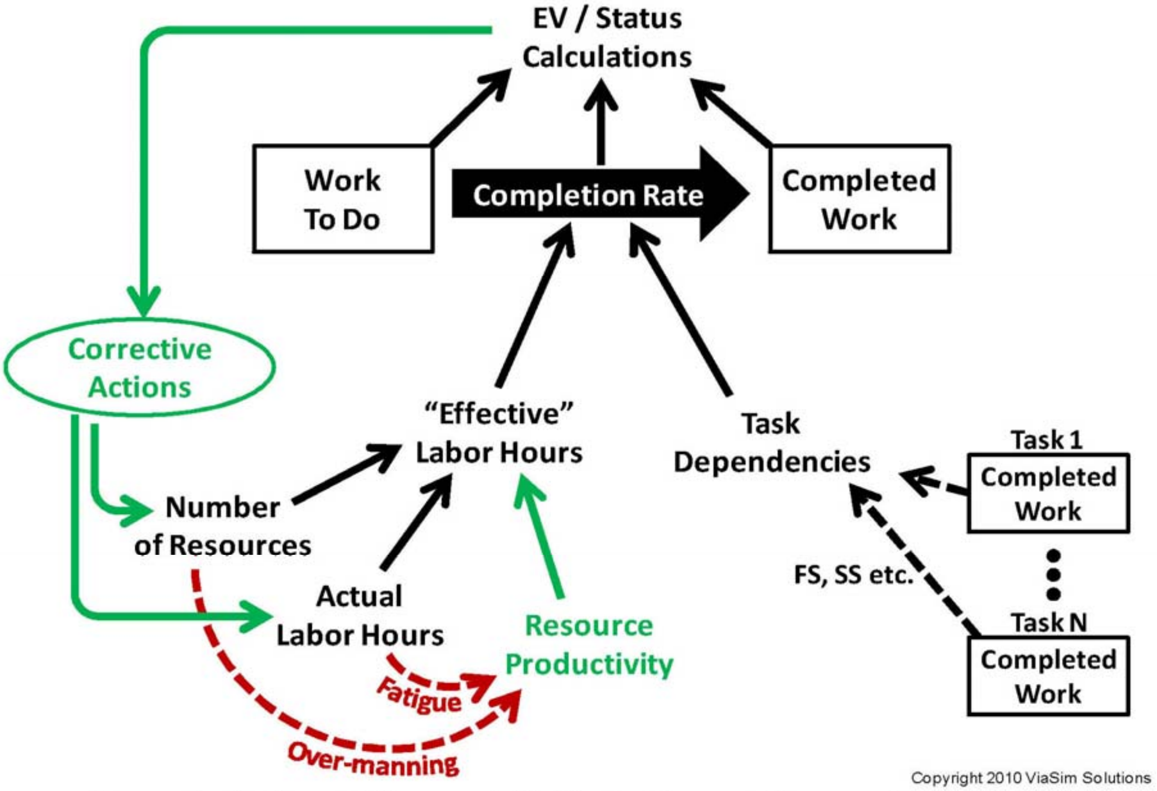


Figure 4: Additional Elements Added by the Dynamic Progress Method (DPM)

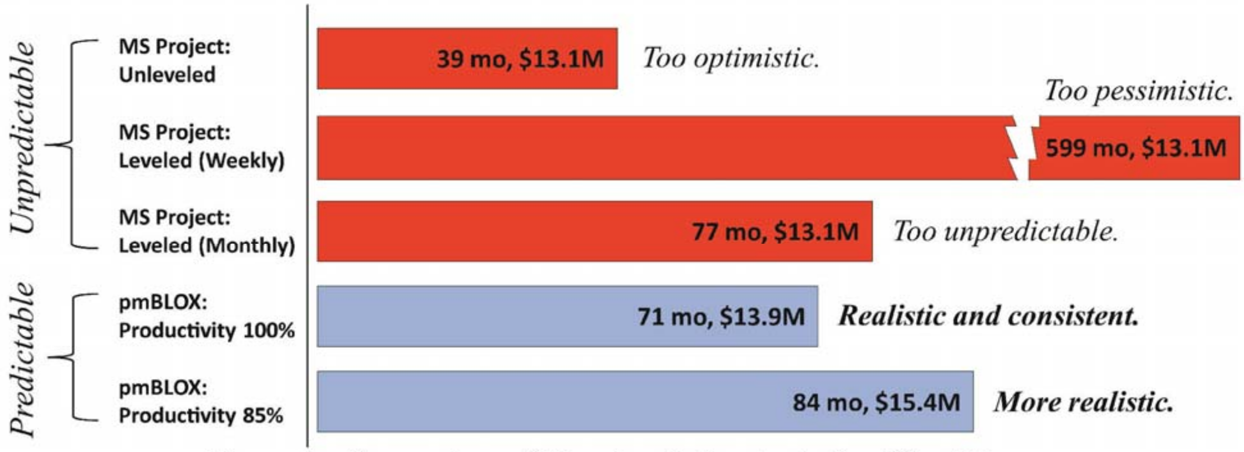


Figure 5: Comparison of Planning Estimates in Bar-Chart Form

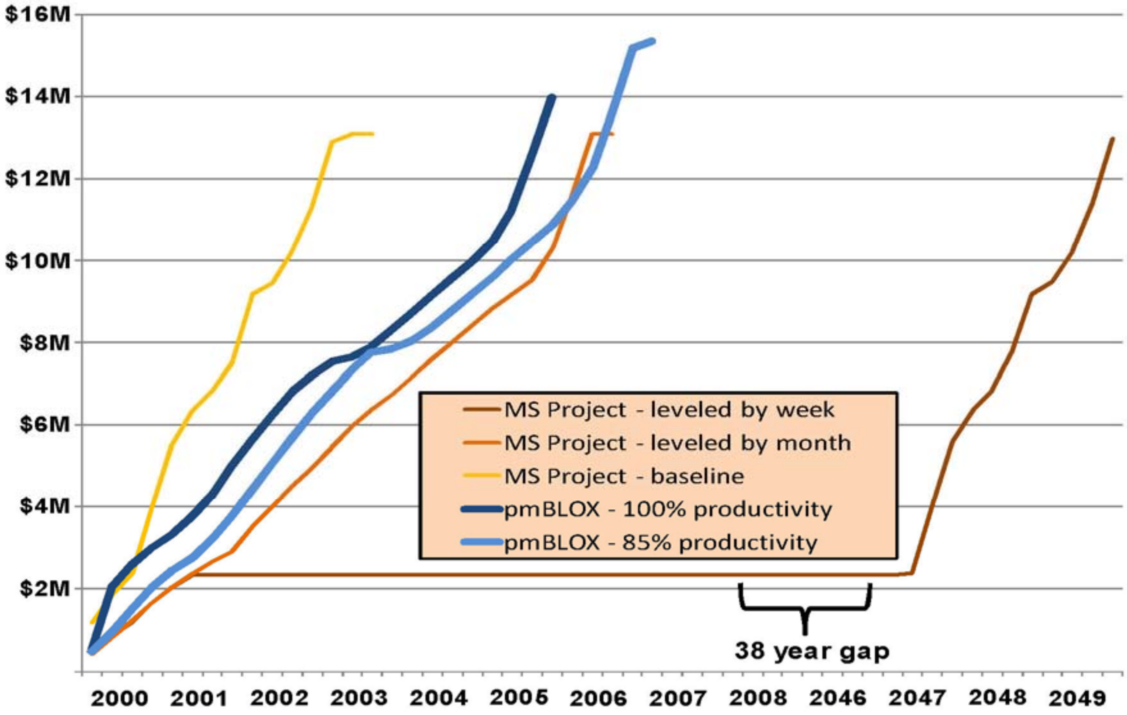
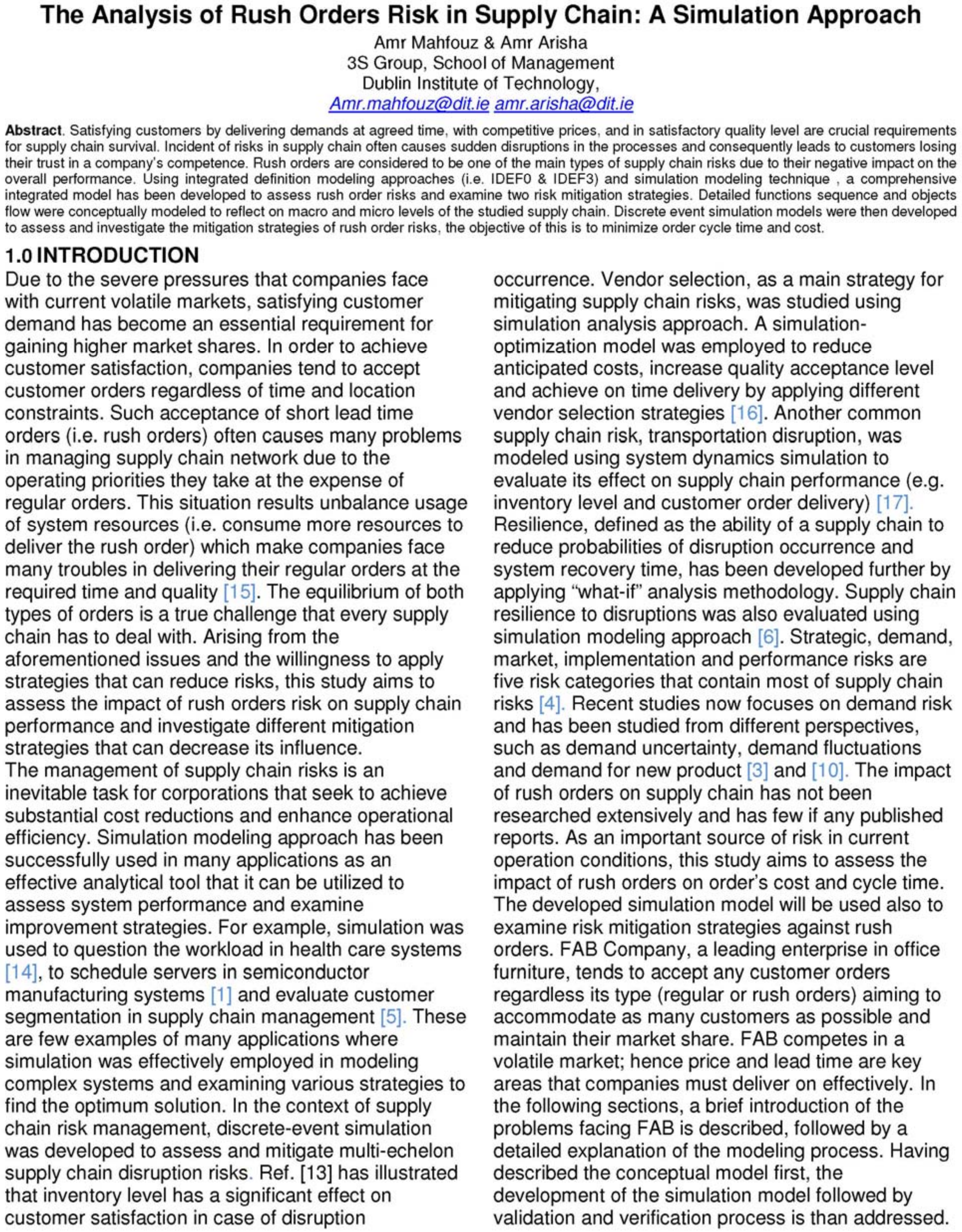


Figure 6: Comparison of Planning Estimates in Line-Chart Form

2.6 The Analysis of Rush Orders Risk in Supply Chain: A Simulation Approach



The Analysis of Rush Orders Risk in Supply Chain: A Simulation Approach

Amr Mahfouz & Amr Arisha
3S Group, School of Management
Dublin Institute of Technology,
Amr.mahfouz@dit.ie amr.arisha@dit.ie

Abstract. Satisfying customers by delivering demands at agreed time, with competitive prices, and in satisfactory quality level are crucial requirements for supply chain survival. Incident of risks in supply chain often causes sudden disruptions in the processes and consequently leads to customers losing their trust in a company's competence. Rush orders are considered to be one of the main types of supply chain risks due to their negative impact on the overall performance. Using integrated definition modeling approaches (i.e. IDEF0 & IDEF3) and simulation modeling technique , a comprehensive integrated model has been developed to assess rush order risks and examine two risk mitigation strategies. Detailed functions sequence and objects flow were conceptually modeled to reflect on macro and micro levels of the studied supply chain. Discrete event simulation models were then developed to assess and investigate the mitigation strategies of rush order risks, the objective of this is to minimize order cycle time and cost.

1.0 INTRODUCTION

Due to the severe pressures that companies face with current volatile markets, satisfying customer demand has become an essential requirement for gaining higher market shares. In order to achieve customer satisfaction, companies tend to accept customer orders regardless of time and location constraints. Such acceptance of short lead time orders (i.e. rush orders) often causes many problems in managing supply chain network due to the operating priorities they take at the expense of regular orders. This situation results unbalance usage of system resources (i.e. consume more resources to deliver the rush order) which make companies face many troubles in delivering their regular orders at the required time and quality [15]. The equilibrium of both types of orders is a true challenge that every supply chain has to deal with. Arising from the aforementioned issues and the willingness to apply strategies that can reduce risks, this study aims to assess the impact of rush orders risk on supply chain performance and investigate different mitigation strategies that can decrease its influence.

The management of supply chain risks is an inevitable task for corporations that seek to achieve substantial cost reductions and enhance operational efficiency. Simulation modeling approach has been successfully used in many applications as an effective analytical tool that it can be utilized to assess system performance and examine improvement strategies. For example, simulation was used to question the workload in health care systems [14], to schedule servers in semiconductor manufacturing systems [1] and evaluate customer segmentation in supply chain management [5]. These are few examples of many applications where simulation was effectively employed in modeling complex systems and examining various strategies to find the optimum solution. In the context of supply chain risk management, discrete-event simulation was developed to assess and mitigate multi-echelon supply chain disruption risks. Ref. [13] has illustrated that inventory level has a significant effect on customer satisfaction in case of disruption occurrence. Vendor selection, as a main strategy for mitigating supply chain risks, was studied using simulation analysis approach. A simulation-optimization model was employed to reduce anticipated costs, increase quality acceptance level and achieve on time delivery by applying different vendor selection strategies [16]. Another common supply chain risk, transportation disruption, was modeled using system dynamics simulation to evaluate its effect on supply chain performance (e.g. inventory level and customer order delivery) [17]. Resilience, defined as the ability of a supply chain to reduce probabilities of disruption occurrence and system recovery time, has been developed further by applying "what-if" analysis methodology. Supply chain resilience to disruptions was also evaluated using simulation modeling approach [6]. Strategic, demand, market, implementation and performance risks are five risk categories that contain most of supply chain risks [4]. Recent studies now focuses on demand risk and has been studied from different perspectives, such as demand uncertainty, demand fluctuations and demand for new product [3] and [10]. The impact of rush orders on supply chain has not been researched extensively and has few if any published reports. As an important source of risk in current operation conditions, this study aims to assess the impact of rush orders on order's cost and cycle time. The developed simulation model will be used also to examine risk mitigation strategies against rush orders. FAB Company, a leading enterprise in office furniture, tends to accept any customer orders regardless its type (regular or rush orders) aiming to accommodate as many customers as possible and maintain their market share. FAB competes in a volatile market; hence price and lead time are key areas that companies must deliver on effectively. In the following sections, a brief introduction of the problems facing FAB is described, followed by a detailed explanation of the modeling process. Having described the conceptual model first, the development of the simulation model followed by validation and verification process is than addressed.

Finally, results and areas of discussion are presented.

2.0 PROBLEM STATEMENT

FAB supplies various products such as, staff desks, workstations, chairs, doors, curtain walls, floors, partitions, ceiling, cladding and many more products (i.e. over 2000 products). Along with these products, complete furnishing solutions are offered including; bank sectors, entertainment facilities (i.e. cinemas), tourism buildings (i.e. hotels) and the health care sector (i.e. hospitals). The company is a typical example of a multi-echelon complex supply chain (Fig. 1). Customer orders arrive in the form of individual contracts; each contract usually contains different product types with different quantities. Contracts are dispatched to singular products and parts which are designed by the design office (i.e. design tier). After that, products and parts designs are sent to research and development (R&D) to create the Bill of Materials (BOMs). Both drawings and BOMs are passed to the planning for; (1) checking the availability of in-stock raw materials, (2) issuing purchase orders for unavailable raw materials, (3) assigning factories' schedule to allocate required work orders. Once raw materials are available, manufacturing processes then start. Distribution centre staff then collects the final products from factories, which are packed and shipped to the end customer. The intense competition that FAB faces in today's market in addition to its desire to compete on a worldwide scale prompts them to shape their supply chain in Engineer-To-Order (ETO) structure and to accept both regular and rush orders. ETO structure usually increases the level of complexity in the relationship between the various entities of supply chain [7]. However it has to be used to create competitive leverage by producing customized products according to customer specifications (based on customer site layout and characteristics). On the other hand, accepting rush orders protects the company market share by maintaining customer satisfaction which in turn creates a wider niche and extra profit margins. Nevertheless, rush orders might cause delay in delivering regular orders as they have priority in using available resources and raw materials. Given the complexity of FAB's supply chain and the large number of variables, entities and operations rules included, there is a great need to use an effective methodology to support FAB decision makers. Simulation modeling is used as a powerful tool that can handle such complexity and be used as an efficient risk assessment tool. FAB simulation models are developed with three different scenarios in order to assess rush order risk and to investigate different mitigation strategies to reduce risk's influence. Average cycle times and costs are the two main performance measures that are used in this study to achieve the following objectives:

(1) Develop a simulation model for the FAB supply chain in order to model ETO complexity.

(2) Use simulation models to assess the impact of rush orders on system performance.

(3) Investigate new risk mitigation strategies against the current strategies and select the best regarding to the performance indicators.

3.0 RESEARCH METHODOLOGY

The research methodology encompasses various phases starting with adopting conceptual models, data collection, developing simulation models, validation and verification and ending with results analysis and discussion.

3.1. Conceptual Model

Integration definition model (IDEF) family has been used to conceptually model the FAB supply chain. Hierarchical structure of IDEF language allows users (e.g. strategic managers, operational engineers and system analyzers) to effectively understand the sequence and details of system's functions. IDEF

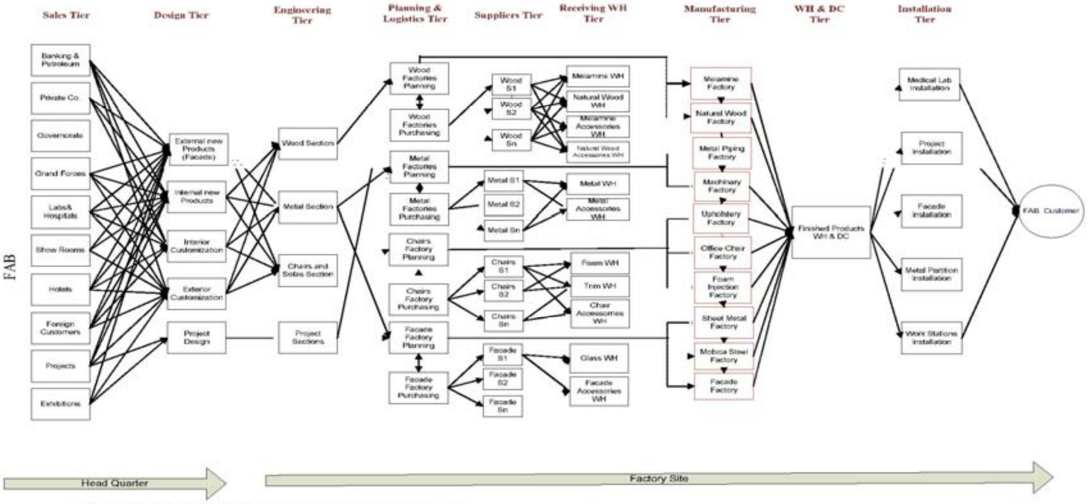


Figure 1 FAB Supply Chain structure

language has different kinds of structures that can model systems with various purposes [12]. IDEF0 (i.e. functional modeling) and IDEF3 (i.e. process modeling) are most relative techniques for business process modeling task. The main difference between both techniques is that IDEF0 focuses more on how business functions are defined, sequenced and connected by their inputs and resources. On the other hand, IDEF3 is a more detailed modeling approach that represents the logical object's flow through system's processes [9]. In this study, IDEF0 and IDEF3 are integrated in order to model the complexity of FAB supply chain. The macro-level is modeled using IDEF0 showing the functions within FAB, their inputs, outputs, controls and resources (Fig. 3). On the other hand, IDEF3 is used to model the micro-level of FAB focusing on products flow and system's operational rules (Fig.4).

3.2. Modeling Macro-level Using IDEF0

An activity block, which is the main unit of IDEF0, describes the main functions of FAB supply chain. Inputs, outputs, mechanisms, and controls are represented by horizontal and vertical arrows (Fig. 2). In addition to input and output arrows, the mechanisms arrow shows resources that facilitate modeled functions (e.g. labors, machines, computer systems, etc...). Function control arrows (top arrow) can be company regulations, standards or legislation. Different kinds of orders are received by FAB's sales staff. Some of these orders, like projects and tenders, require a preparation of time plans and financial offers before issuing the contracts. The ordered products are classified into two main categories; Standard Products (products that were manufactured in the company before and all its details are available) and Special Products (products with new designs and specifications and no data available for them). Standard Products go directly to the planning process in order to supply required materials and to issue factories' work orders. On the other hand, an engineering process starts by joining sales, design and R&D staff to design and identify Special Product's BOM. A detailed BOM is passed to the planning process again to supply required materials and issue work orders. After the production process, all products are collected according to their contract numbers and then delivered and installed at customer sites. Figure 3 graphically shows the main sequence of FAB supply chain processes, their inputs, outputs, and resources.

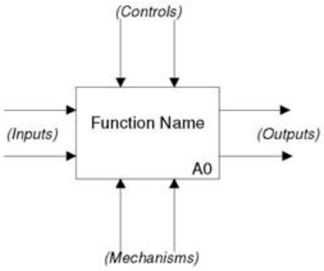


Figure 2 Basic IDEF0 construct

3.3. Modeling Micro-level Using IDEF3

In contrast to IDEF0, IDEF3 has less strict syntax and semantic rules. Integrating IDEF0 and IDEF3 using the hierarchical structure provides a detailed conceptual model for FAB. For example, Fig.(4) shows that the design and development function (A2) was split to more elaborated processes representing the main flow of design& development function. IDEF3 acts as a bridge between general conceptual model of FAB and proposed simulation model. In simulation model, processes blocks that are shown in IDEF3 model is considered as an activity blocks, whereas branching points will be represented as routing decisions.

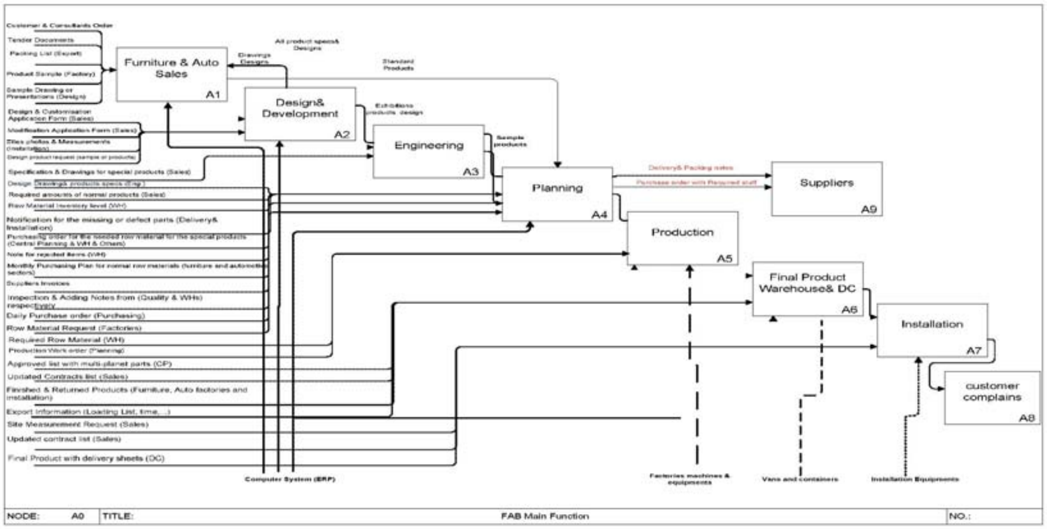

Customer & Consultants Order
Tender Documents
Packing List (Export)
Product Sample (Factory)
Sample Drawing or Presentations (Design)
Design & Customization Application Form (Sales)
Modification Application Form (Sales)
Sites photos & Measurements (Installation)
Design product request (sample or products)
Specification & Drawings for special products (Sales)
Design Drawings& products specs (Eng.)
Required amounts of normal products (Sales)
Raw Material Inventory level (WH)
Notification for the missing or defect parts (Delivery& Installation)
Purchasing order for the needed row material for the special products (Central Planning & WH & Others)
Note for rejected items (WH)
Monthly Purchasing Plan for normal row materials (furniture and automotive sectors)
Suppliers Invoices
Inspection & Adding Notes from (Quality & WHs) respectively
Daily Purchase order (Purchasing)
Row Material Request (Factories)
Required Row Material (WH)
Production Work order (Planning)
Approved list with multi-planet parts (CP)
Updated Contracts list (Sales)
Finished & Returned Products (Furniture, Auto factories and installation)
Export Information (Loading List, time,...)
Site Measurement Request (Sales)
Updated contract list (Sales)
Final Product with delivery sheets (DC)
Furniture & Auto Sales
A1
Drawings Designs
All product specs& Designs
Standard Products
Design& Development
A2
Exhibitions products design
Engineering
A3
Sample products
Planning
A4
Delivery& Packing notes
Purchase order with Required stuff
Suppliers
A9
Production
A5
Final Product Warehouse& DC
A6
Installation
A7
customer complains
A8
Computer System (ERP)
Factories machines & equipments
Vans and containers
Installation Equipments
NODE:
A0
TITLE:
FAB Main Function
NO.:

Figure 3 IDEF0 model for FAB supply chain

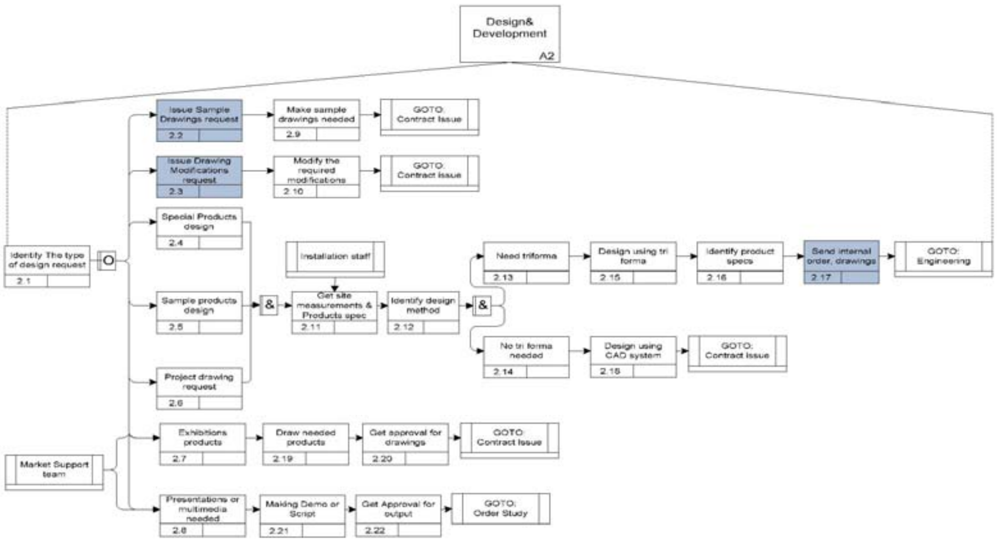


Figure 4 IDEF3 model for FAB supply chain

4.0SIMULATION MODEL

Conceptual models (Fig. 3 and 4) provide a clear understanding of the relationships between different system entities. They also show the resources that are required at each step. Discrete-event simulation was employed to develop a detailed simulation model to mimic the process of FAB. Model assumptions are (i) no supplier disruptions are considered (ii) all received materials are accepted (no return of poor quality materials). The model was built and executed using simulation software based on Java and XML technology which provides object-oriented hierarchical and event-driven simulation capabilities for modeling large-scale applications. It also utilizes breakthrough activity-based modeling paradigms (e.g. real world activities such as assembly, batching, and branching). Many features in FAB are coded in the simulated blocks to mimic the real life application characteristics.

System entities are the objects (products) that are modified by resources (sales staff, design staff, supplier staff, etc...). Resources are characterized by their availability, whereas the product entity is characterized by arrival time, processing time, and product characteristics. Logical entities make decisions for creating, joining, splitting, buffering, and branching product entities. The model contains 950 blocks representing; queues, activities, and branching points. The hierarchical feature has been used to mimic exactly the system flow. Two main layers are developed; (1) the upper layer (i.e. macro level) which represents the main activities in FAB. Five main activities cover all simulated processes contain; sales and customer support, products design and engineering, resource planning and material management, manufacturing, and finally the warehouse and installation activity. These activities contain 13 processes that are conceptually modeled by IDEF0 at Fig. (3) representing the core processes of FAB supply chain such as sales, contract issue, design, engineering, material management, …, etc. (2) The lower layer (micro-level) illustrates the objects' flow of each process. IDEF3 was used to develop the conceptual model for this layer, where all operations rules are represented. An example of the lower layer modeling that illustrates inputs, outputs, and the relationships between some upper level processes is shown in Fig. (5). For the model to reach its steady state condition, the warm-up was 100 hours. Every simulation run represented a year of actual timing. In the experimental phase, the average from 10 replications of average cycle time and average operations costs were used as the main performance measures.

Table 1 shows the main input variables of FAB supply chain. Theoretical statistical distribution was utilized to represent the random patterns of input variables. The analysis of demand data resulted in normal distribution with a mean of 4 days and standard deviation of 1 day. Consistent with what was reported by [2], service time was fitted to exponential distribution since service time data are completely random. For production time weibull distribution was used. Finally, due to the shortage of lead time data that were supplied by FAB, gamma distribution was used, according to [8].

Table 1: Model input variables

Category	Input Variable
Customers	Orders arrivals
Suppliers	• Lead time • Incoming inspection time
Plants	Production time per product
Other Processes	Service time
Transportation	Time to ship between plant and head quarter

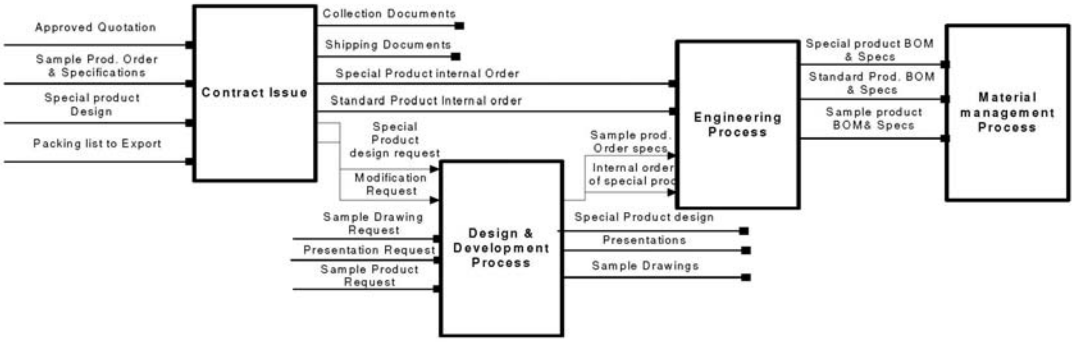


Figure 5 Example of micro level of the simulation model

4.1. Model Validation and Verification

In an effort to make the decisions that are based on simulation models more accurate, efficient methods of verification and validation (V&V) are needed. Inaccurate simulation results always lead to wrong decisions proposals and implementation, resulting in high costs that can be more than the total cost used for the simulation study. Therefore, the correctness and suitability of simulation results are very important. Different methods are used in order to verify simulation coding. Decomposition method (i.e. verify every group of blocks) was used to insure that every block functions as expected. A built-in simulation debugger is also used to avoid any coding bugs. On the other hand, validation process was considered as an integral process, which starts from input data collection through conceptual and simulation model development and ends at output data analysis. Out of 10 V&V methods that have been mentioned in [11], three validation methods have applied in three phases of this study; (1) data collection phase, (2) conceptual modeling phase and finally (3) simulation results phase. Three main objectives were targeted in the validation process of data collection phase; (1) no measurement errors in data collection process, (2) generated data have to match the pattern of historical data and (3) attribute values are within specified range. To achieve that, a detailed examination of data documentation's quality and consistency was done with the cooperation of FAB company staff. In addition, real data were compared with statistically generated data and results were approved by the IT department. The conceptual model was validated based on structured interviews with system managers and staff in order to be certain that all specified processes, structures, system elements, inputs and outputs are considered correctly. The modeling team also examined the accuracy and consistency of the conceptual model to the problem definition. After that, system performance indicators were revised with decision makers in order to be sure that it fits model objectives. Finally, two main approaches were used to validate the final simulation results. The first is "Face validation" approach that was performed by interviewing managers and manufacturing teams in order to validate simulation results. The second is "comparison test", which is achieved by comparing the model and system output under identical input conditions. The validation process has shown that there is only 15% deviation between simulated and actual results.

5.0 RISK ANALSYIS

In this paper, risk analysis procedure contains two main phases. First, it focuses on assessing the influence of rush order risk on performance indicators. Second, risk mitigation strategies have been examined against current system configuration. Three simulation models were developed representing three operating strategies of FAB supply chain;

(1) First Strategy (no rush order strategy): represents the current system configurations for FAB Company. No rush orders received, is the only assumption in this strategy.

(2) Second Strategy (mixed orders strategy): represents current system configuration for FAB Company. Both regular and rush orders are expected in this strategy. The two orders' types have the same route through supply chain, however rush orders have higher priority for using resources and raw materials over regular orders.

(3) Third Strategy (independent production route strategy): represents risk mitigation strategy which suggests a separate processing route for both rush and regular orders with dedicated resources for each of them. Ten simulation runs for no rush order and mixed orders strategies are illustrated at Table 2 and 3.

Table 2: Simulation results of no-rush order strategy

Strategy	Number of Replications	Orders Types				
		Regular Orders			Rush Orders	
		Cycle Time (days)	Cost (euro)		Cycle Time (days)	Cost (euro)
No rush orders strategy	1	31.44	7959.68		0	0
	2	25.66	7384.82		0	0
	3	29.32	7747.05		0	0
	4	29.84	7813.07		0	0
	5	29.10	8618.53		0	0
	6	27.23	7816.21		0	0
	7	26.81	7671		0	0
	8	28.00	7702.03		0	0
	9	38.87	8176.16		0	0
	10	30.85	8116.38		0	0

Table 3: Simulation results of mixed-order strategy

Strategy	Number of Replications	Orders Types				
		Regular Orders			Rush Orders	
		Cycle Time	Average Cost		Cycle Time	Average Cost
Mixed orders strategy	1	48.40	11074.44		22.22	9549.57
	2	37.00	9180.82		24.03	9134.61
	3	29.96	10433.74		21.44	7550
	4	35.61	10234.68		24.51	8511.3
	5	42.96	11076.76		27.38	9707.05
	6	31.83	10153.27		21.82	7353.75
	7	54.00	10856.41		23.78	8308.36
	8	61.75	11166.25		23.21	8225.44
	9	71.02	10495.88		27.66	9483.19
	10	37.65	9570.09		27.53	9352.08

Performance indicators were divided into two sections (regular and rush orders). No results are reported for rush order's columns in table 2 as no rush orders were allowed in this strategy. For the no rush order strategy, average cycle time of regular orders was ranged between 25 to 31 days, whereas average cost varies between € 7300 and € 8200 per order. Time and cost figures of regular orders were increased in the mixed orders strategy to record average cycle time between 30 to 70 days and average cost between € 9000 and € 11000 (Table 3). Table 4 shows the differences between the two strategies and indicates an increasing in average cycle time and average cost by 35% and 25% respectively in case of applying mixed orders strategy. Out of these results, it can be concluded that receiving rush orders increases the values of average cycle time and average cost for regular orders. These negative results were generated as a result of the high priority that rush orders take over regular orders along all supply chain processes. Long waiting time, process interruptions and resources unavailability are the challenges that face regular orders in case of receiving rush orders. Inaccurate delivery time with high prices for regular orders are the results of rush orders risks under the current system configuration. In order to mitigate the influence of rush order's risk, a separate route strategy (strategy 3) is applied at design, engineering, planning, purchasing, production, and distribution centers. The simulation model was developed for risk mitigation strategy and was investigated against mixed orders strategy. Table 5 recorded a reduction in regular order's delivery time and average cost compared to mixed orders strategy (Table 3). Delivery time was alternated between 25 days to 34 days whereas average cost ranged between €7000 and €9000. The separate route strategy achieved a 37% reduction for regular order's delivery time and about a 17% reduction in average cost (Table 6). Results improvement has been achieved by separating the flow of rush and regular orders' causing a reduction in regular order's waiting time and cost. On the other hand, a reduction in rush order's cycle time and cost by 3% and 27% was noticed in case of applying separate route strategy. Applying the separate processing route strategy did not make a significant impact on rush orders' average cycle time, while average cost was significantly influenced. This can be explained that rush orders in both strategies did not stay in the processes' buffers for a long time. In the mixed order strategy, rush orders were located in the head of all buffers due to the high priority they have over the regular orders. Whereas in separate processing route strategy, orders are split into two processing flows causes a decrease in resources' utilization and hence declined the time of waiting free resources.

Table 4: Differences of performance indicators in applying first and second strategy

		Studied Strategies		Difference between 1st and 2nd Strategy
Order Type	Performance Indicator	No Rush Order Strategy	Mixed Order Strategy	Increased
Regular Order	Cycle Time	29.716	45.025	35%
	Average Cost	7900.493	10424.234	25%
Rush Order	Cycle Time	0	24.362	100%
	Average Cost	0	8717.535	100%

Table 5: Simulation results of separate route strategy

Strategy	Number of Replications	Orders Types				
		Regular Orders			Rush Orders	
		Cycle Time	Average Cost		Cycle Time	Average Cost
Separate route strategy	1	34.00	8738.82		26.11	6799.42
	2	30.30	8866.41		25.22	6849.81
	3	27.37	8406.08		24.39	6520.65
	4	26.04	7600.91		22.99	6361.09
	5	26.02	8560.93		23.22	5988.47
	6	25.80	9006.65		22.39	5699.97
	7	31.95	9082.7		23.53	5764.3
	8	26.66	8423.42		23.80	6363.76
	9	28.28	8347.94		24.50	6805.84
	10	25.98	9149.19		23.06	6131.45

Table 6: Differences of performance indicators in applying second and third strategies

		Studied Strategies		Difference between 2nd and 3rd Strategy
Order Type	Performance Indicator	Mixed orders strategy	Separate route strategies	Decreasing
Regular Order	Cycle Time	45.025	28.245	37%
	Average Cost	10424.234	8618.305	17%
Rush Order	Cycle Time	24.362	23.926004	3%
	Average Cost	8717.535	6328.476	27%

6.0 CONCLUSION

Rush orders is a challenging risk for supply chains due to their nature of pre-emption over regular orders processing. Due to severe competition in current markets, enterprises have no longer an option not to consent rush orders, even with such inconvenient operating conditions that might include the restructuring of supply chain strategy. This often leads to the adoption of complex structures such as Engineer-To-Order (ETO). Simulation modeling has proven to be an effective tool to handle systems featuring high levels of complexity with uncertainty. Hence, simulation models were developed in order to

effectively assess the impact of rush orders risks on system performance indicators (cycle time and total cost). It was also used to investigate risk mitigation strategy that can decrease the negative impact of rush orders on FAB's supply chain performance. IDEF0 and IDEF3 were integrated to develop a detailed conceptual model of FAB's supply chain. IDEF language was used as it applies a standard format with hierarchical structure that supports the modeling of predecessors, relationships, inter-relationship and interdependences of activities and objects. Modeling has been structured into layers to be able to present the processes and their activities as well as the overall system view. Three methods of validation were applied for data collection phase, conceptual modeling phase and then simulation modeling results.
To assess and mitigate the impact of rush orders risk, this study focused on three strategies (i.e. no rush order strategy, mixed orders strategy, and independent route strategy). The simulation model provides not only numerical measures of system performance, but also insights about the effect of rush orders on the delivery time and the cost of regular orders. Results showed that rush orders have a negative impact on both cycle time and average cost of regular orders as they were increased by (35%) and (25%) respectively. The risk mitigation strategy - dedicate a separate route for rush orders-minimized the impact of rush orders by decreasing cycle time by 37% and cost by 17% for regular orders and 3% and 27% for rush orders.

7.0 REFERENCES

1- Arisha, A. and Young, P. (2004). “Intelligent simulation-based lot scheduling of photolithography toolsets in a wafer fabrication facility”. *Proceedings of the 2004 winter simulation conference.*

2- Banks, J., Carson, J., Nelson, B. and Nicol, D. (2005). “Discrete Event System Simulation”. 4/e: Pearson Prentice Hall, Upper Saddle River, NJ 07458.

3- Chen, X. and Zhang, J. (2008). “Supply chain risks analysis by using jump-diffusion model”. *Proceedings of the 2008 winter simulation conference.*

4- Clouse, M. and Busch, J. (2003). “How to Identify and Manage Supply Risk”. Supply Chain Planet October.

5- Chen, Y. and Xueping, L. (2009). “The effect of customer segmentation on an inventory system in the presence of supply disruptions”. *Proceedings of the 2009 winter simulation conference.*

6- Falasca, M., Zobel C. and Cook. D. (2008).“A decision support framework to assess supply chain resilience”. *Proceedings of the 5th international ISCRAM conference-Washigton, SC, USA.*

7- Gosling, J., and Naim, M. M. (2009). "Engineer-to-order supply chain management: A literature review and research agenda". *International journal of production economics 122: 741-754.*

8- Haddley, G. and Whitin, T. (1963). “Analysis of Inventory System”. *New Jersey: Prentice Hall*

9- Mahfouz, A., Ali, S. and Arisha, A. (2010). “Practical simulation application: evaluation of process control parameters in twisted-pair cables manufacturing system”. *Simulation modeling practice and theory 18: 471-482.*

10- Nagurney, A., Cruz, J., Dong, J. and Zhang, D. (2005). “Supply chain networks, electronic commerce, and supply side and demand side risk”. *European Journal of Operational Research 164: 120-142.*

11 - Rabe, M., Spieckermann, S., and Wenzel, S. (2009). "Verification and validation activities within a new procedure model for V&V in production and logistics simulation”. *Proceedings of the 2009 winter simulation conference.*

12- R.J. Mayer, C.P. Menzel, M.K. Painter, T. Blinn, P.S. Dewitte, Information Integration for Concurrent Engineering (IICE) IDEF3 Process Description Capture Method Report, Knowledge Based Systems Inc., College Station, TX, 1997.

13- Schmitt, J.A. and Singh, M. (2009). “Quantifying supply chain disruption risk using Monte Carlo and discrete-event simulation”. *Proceedings of the 2009 winter simulation conference.*

14- Thorwarth, M., Arisha, A. and Harper, P. (2009). “Simulation model to investigate flexible workload management for healthcare and service escape environment”. *Proceedings of the 2009 winter simulation conference.*

15- Van Donk, D.P. and Van der Vaart, T. (2005).”A case of shared resources, uncertainty and supply chain integration in the process industry”. International Journal of Production Economics 96: 97-108.

16- Wu, D. and Olson, D. (2008). “Supply chain risk, simulation, and vendor selection. *International Journal of Production Economics 114: 646-655.*

17- Wilson, M. (2007). “The impact of transportation disruptions on supply chain performance” *Transportation Research Part E 43: 295-320.*

8.0 Acknowledgment

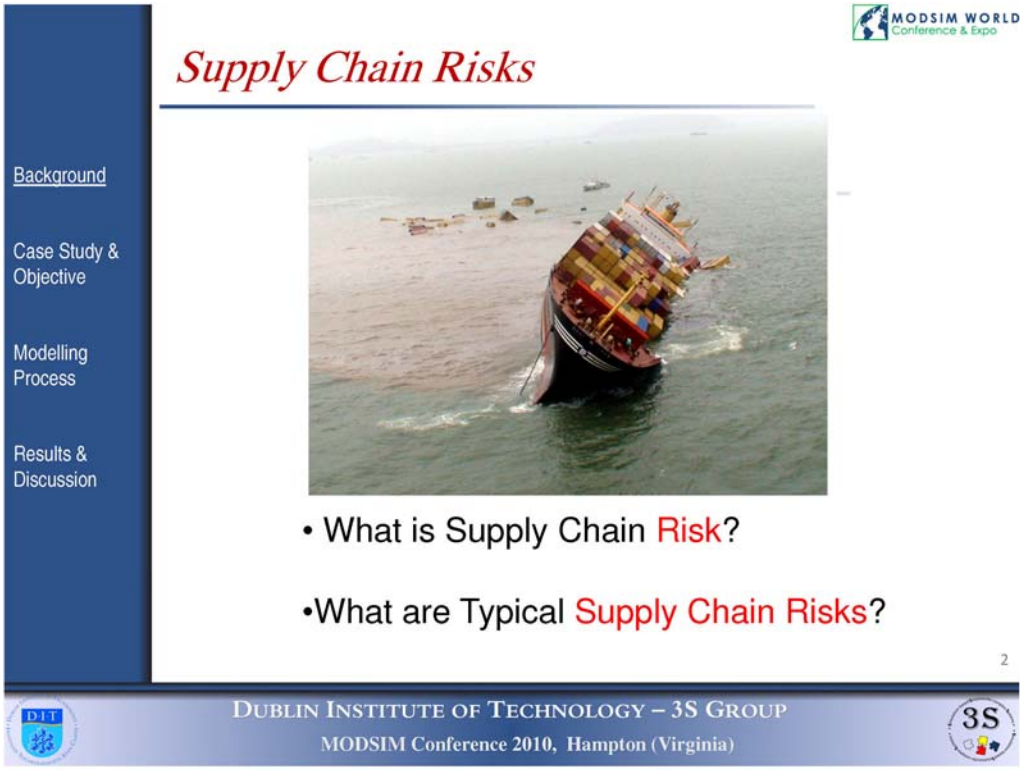
The authors thank the head of business process integration department in FAB Company and all his engineers for their cooperation.

The Analysis of Rush Orders Risk in Supply Chain: A Simulation Approach

Amr Mahfouz

Amr Arisha

3S Group
School of Management
Dublin Institute of Technology
Ireland



Supply Chain Risks

Background

Case Study & Objective

Modelling Process

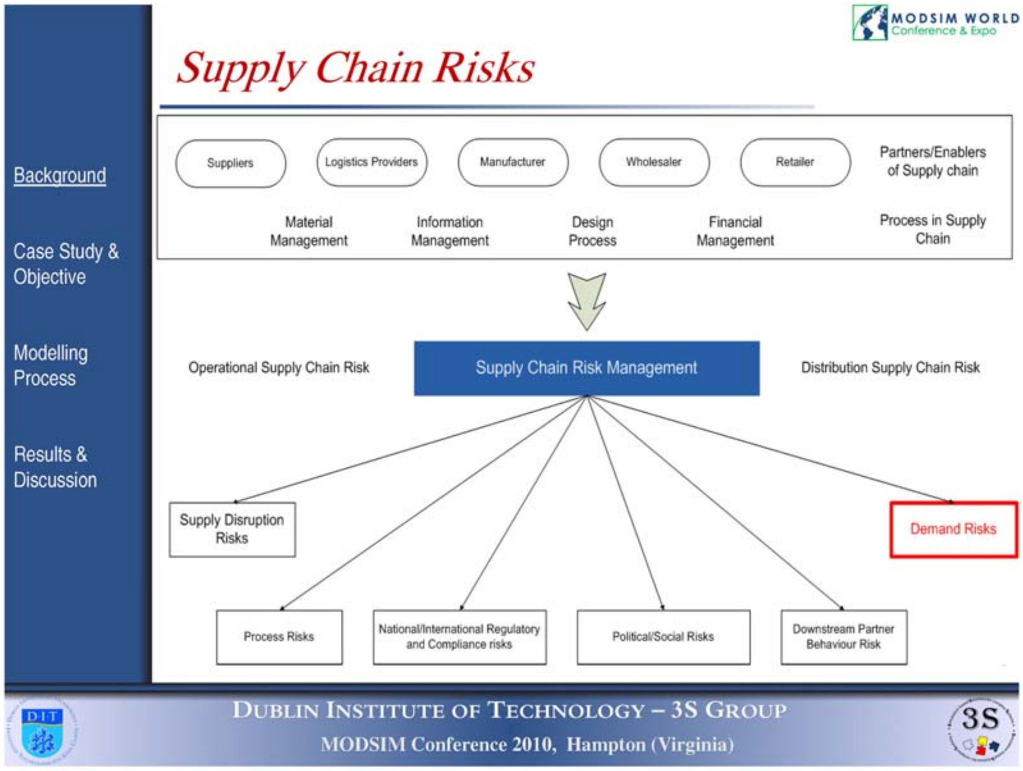
Results & Discussion

- What is Supply Chain Risk?

•What are Typical Supply Chain Risks?

2

DUBLIN INSTITUTE OF TECHNOLOGY – 3S GROUP
MODSIM Conference 2010, Hampton (Virginia)



MODSIM WORLD Conference & Expo

Supply Chain Risks

Background

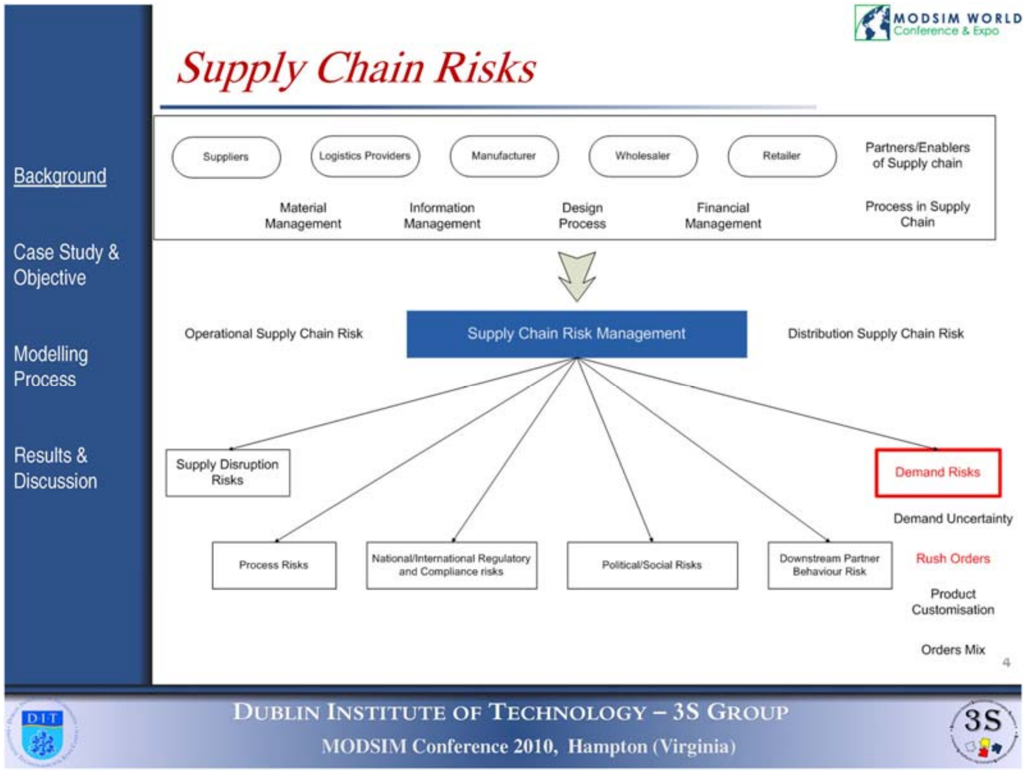
Case Study & Objective

Modelling Process

Results & Discussion

DUBLIN INSTITUTE OF TECHNOLOGY – 3S GROUP

MODSIM Conference 2010, Hampton (Virginia)



MODSIM WORLD Conference & Expo

Supply Chain Risks

Background

Case Study & Objective

Modelling Process

Results & Discussion

Demand Uncertainty

Rush Orders

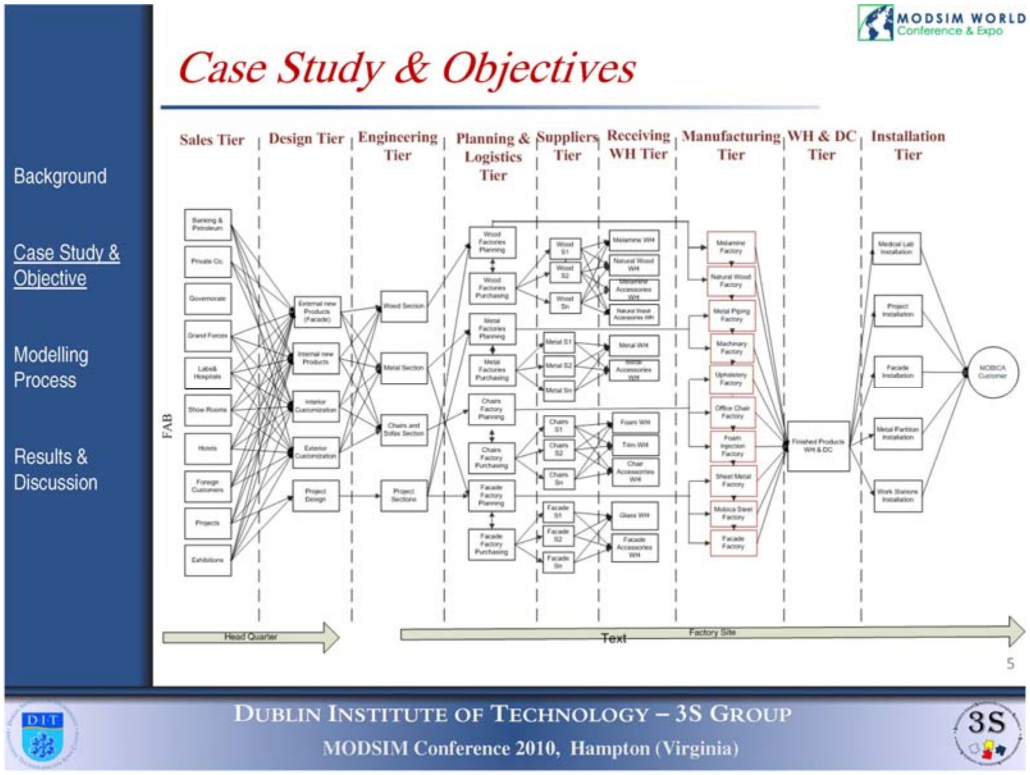
Product Customisation

Orders Mix

4

DUBLIN INSTITUTE OF TECHNOLOGY – 3S GROUP

MODSIM Conference 2010, Hampton (Virginia)



MODSIM WORLD Conference & Expo

Case Study & Objectives

Background

Case Study & Objective

Modelling Process

Results & Discussion

Sales Tier | Design Tier | Engineering Tier | Planning & Logistics Tier | Suppliers Tier | Receiving WH Tier | Manufacturing Tier | WH & DC Tier | Installation Tier

FAB

Banking & Petroleum
Private Co.
Governorate
Grand Forces
Labs& Hospitals
Show Rooms
Hotels
Foreign Customers
Projects
Exhibitions

External new Products (Facade)
Internal new Products
Interior Customization
Exterior Customization
Project Design

Wood Section
Metal Section
Chairs and Sofas Section
Project Sections

Wood Factories Planning
Wood Factories Purchasing
Metal Factories Planning
Metal Factories Purchasing
Chairs Factory Planning
Chairs Factory Purchasing
Facade Factory Planning
Facade Factory Purchasing

Wood S1
Wood S2
Wood Sn
Metal S1
Metal S2
Metal Sn
Chairs S1
Chairs S2
Chairs Sn
Facade S1
Facade S2
Facade Sn

Melamine WH
Natural Wood WH
Melamine Accessories WH
Natural Wood Accessories WH
Metal WH
Metal Accessories WH
Foam WH
Trim WH
Chair Accessories WH
Glass WH
Facade Accessories WH

Melamine Factory
Natural Wood Factory
Metal Piping Factory
Machinery Factory
Upholstery Factory
Office Chair Factory
Foam Injection Factory
Sheet Metal Factory
Mobica Steel Factory
Facade Factory

Finished Products WH & DC

Medical Lab Installation
Project Installation
Facade Installation
Metal Partition Installation
Work Stations Installation

MOBICA Customer

Head Quarter

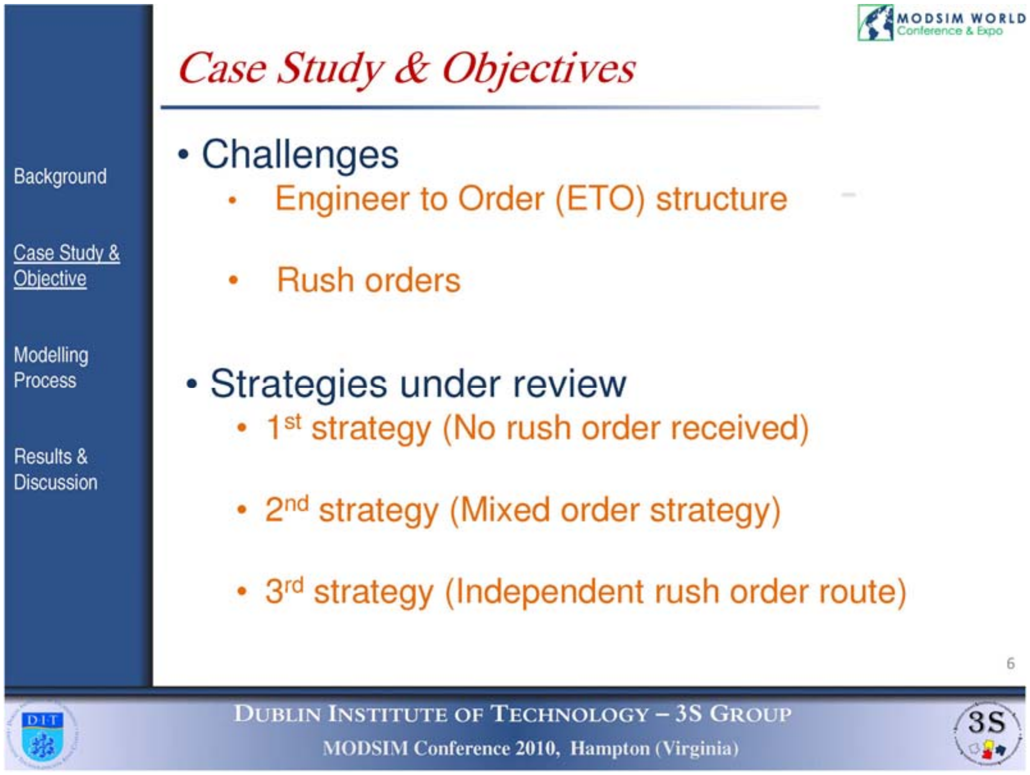
Text

Factory Site

5

DUBLIN INSTITUTE OF TECHNOLOGY – 3S GROUP

MODSIM Conference 2010, Hampton (Virginia)



MODSIM WORLD Conference & Expo

Case Study & Objectives

Background

Case Study & Objective

Modelling Process

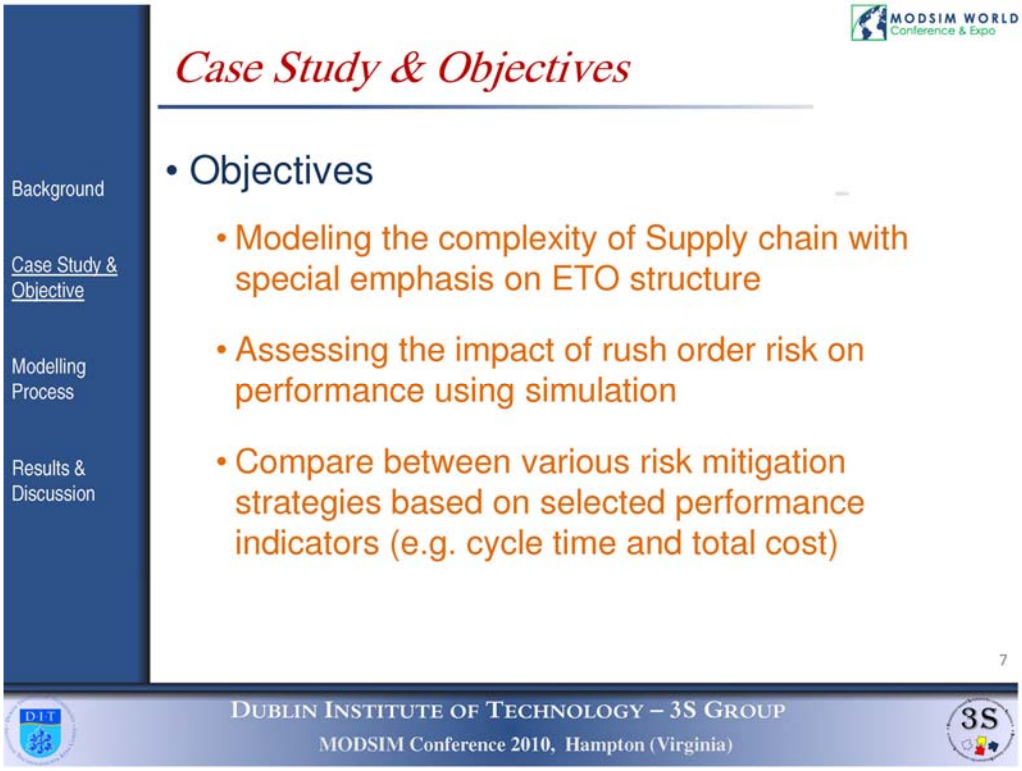
Results & Discussion

- Challenges
  - Engineer to Order (ETO) structure
  - Rush orders
- Strategies under review
  - 1st strategy (No rush order received)
  - 2nd strategy (Mixed order strategy)
  - 3rd strategy (Independent rush order route)

6

DUBLIN INSTITUTE OF TECHNOLOGY – 3S GROUP

MODSIM Conference 2010, Hampton (Virginia)



MODSIM WORLD Conference & Expo

Case Study & Objectives

Background

Case Study & Objective

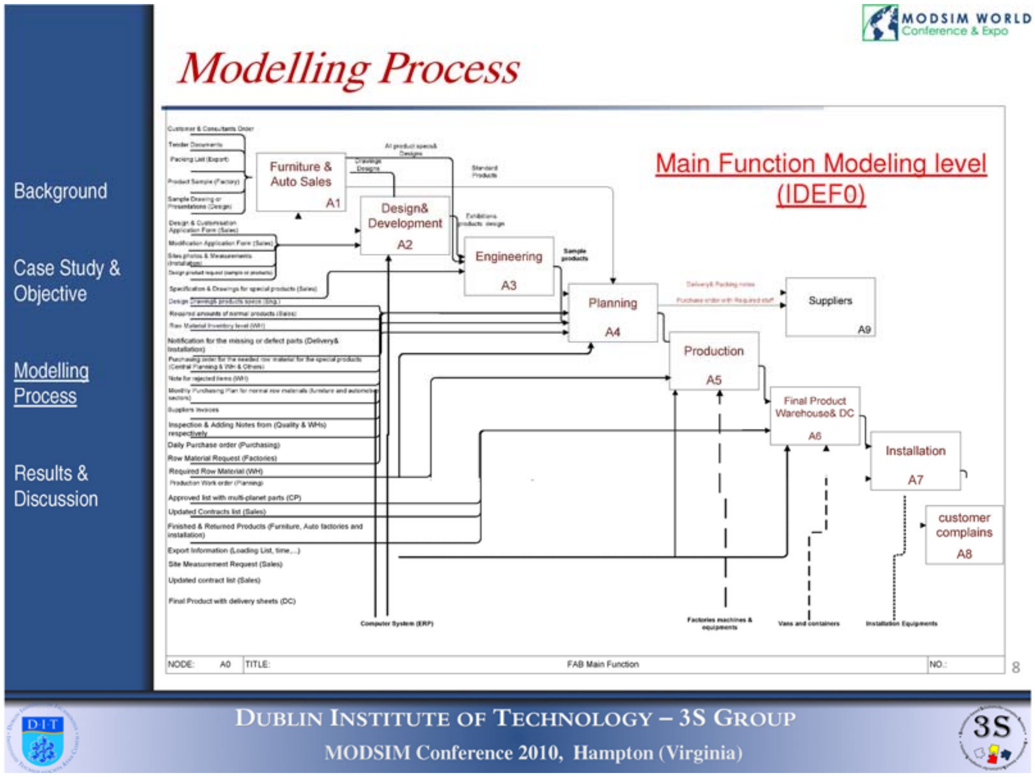
Modelling Process

Results & Discussion

- Objectives
  - Modeling the complexity of Supply chain with special emphasis on ETO structure
  - Assessing the impact of rush order risk on performance using simulation
  - Compare between various risk mitigation strategies based on selected performance indicators (e.g. cycle time and total cost)

7

DUBLIN INSTITUTE OF TECHNOLOGY – 3S GROUP
MODSIM Conference 2010, Hampton (Virginia)



MODSIM WORLD Conference & Expo

Modelling Process

Background

Case Study & Objective

Modelling Process

Results & Discussion

Main Function Modeling level (IDEF0)

Customer & Consultants Order
Tender Documents
Packing List (Export)
Product Sample (Factory)
Sample Drawing or Presentations (Design)
Design & Customisation Application Form (Sales)
Modification Application Form (Sales)
Sites photos & Measurements (Installation)
Design product request (sample or products)
Specification & Drawings for special products (Sales)
Design Drawing& products specs (Eng.)
Required amounts of normal products (Sales)
Raw Material Inventory level (WH)
Notification for the missing or defect parts (Delivery& Installation)
Purchasing order for the needed row material for the special products (Central Planning & WH & Others)
Note for rejected items (WH)
Monthly Purchasing Plan for normal row materials (furniture and automotive sectors)
Suppliers invoices
Inspection & Adding Notes from (Quality & WHs) respectively
Daily Purchase order (Purchasing)
Row Material Request (Factories)
Required Row Material (WH)
Production Work order (Planning)
Approved list with multi-planet parts (CP)
Updated Contracts list (Sales)
Finished & Returned Products (Furniture, Auto factories and installation)
Export Information (Loading List, time,...)
Site Measurement Request (Sales)
Updated contract list (Sales)
Final Product with delivery sheets (DC)

Furniture & Auto Sales A1
Design& Development A2
Engineering A3
Planning A4
Production A5
Final Product Warehouse& DC A6
Installation A7
customer complains A8
Suppliers A9

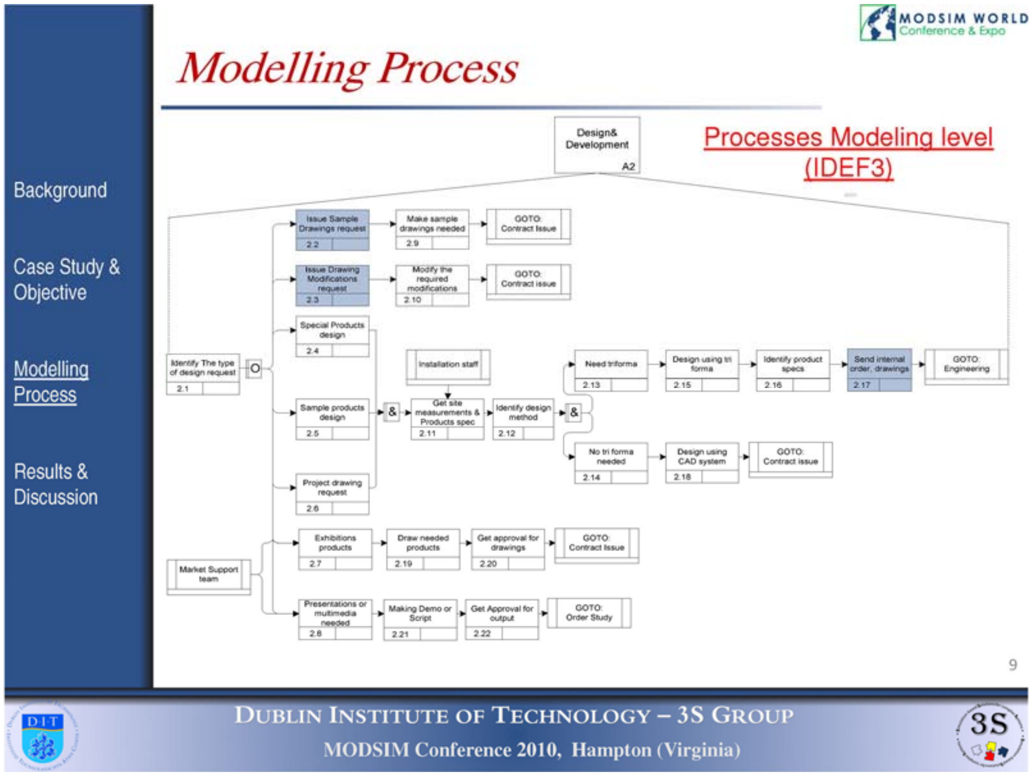
All product specs& Designs
Drawings Designs
Standard Products
Exhibitions products design
Sample products
Delivery& Packing notes
Purchase order with Required stuff

Computer System (ERP)
Factories machines & equipments
Vans and containers
Installation Equipments

NODE: A0 | TITLE: FAB Main Function | NO.:

8

DUBLIN INSTITUTE OF TECHNOLOGY – 3S GROUP
MODSIM Conference 2010, Hampton (Virginia)



MODSIM WORLD Conference & Expo

Modelling Process

Processes Modeling level (IDEF3)

Background

Case Study & Objective

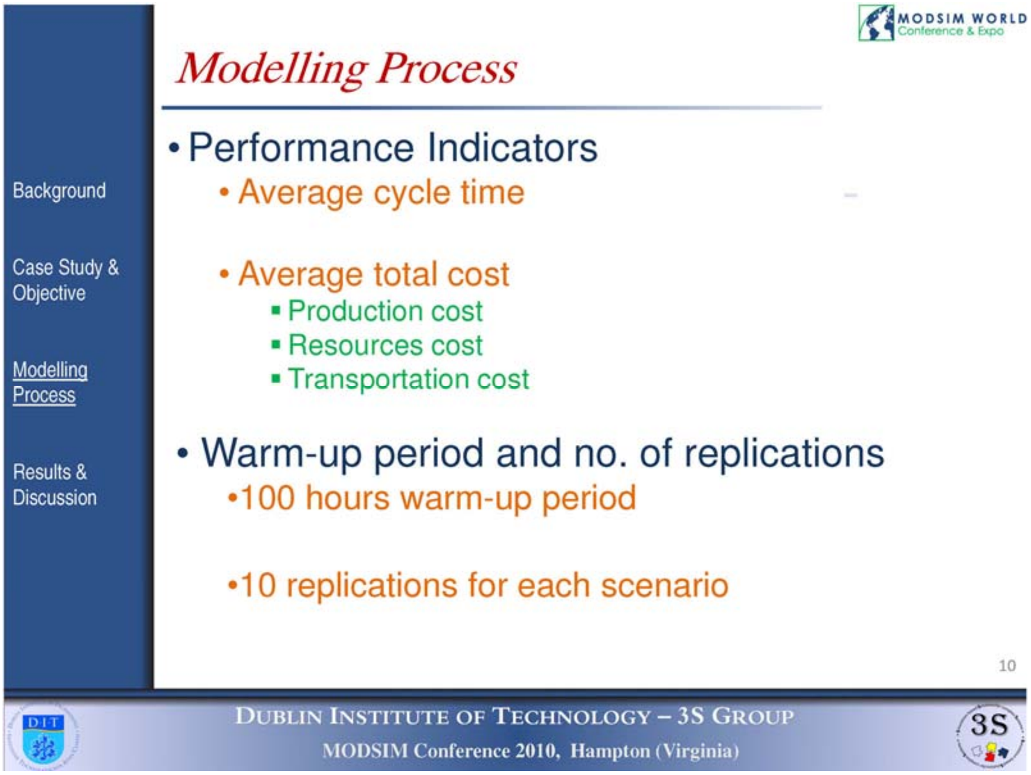
Modelling Process

Results & Discussion

9

DUBLIN INSTITUTE OF TECHNOLOGY – 3S GROUP

MODSIM Conference 2010, Hampton (Virginia)



MODSIM WORLD Conference & Expo

Modelling Process

- Performance Indicators
  - Average cycle time
  - Average total cost
    - Production cost
    - Resources cost
    - Transportation cost
- Warm-up period and no. of replications
  - 100 hours warm-up period
  - 10 replications for each scenario

Background

Case Study & Objective

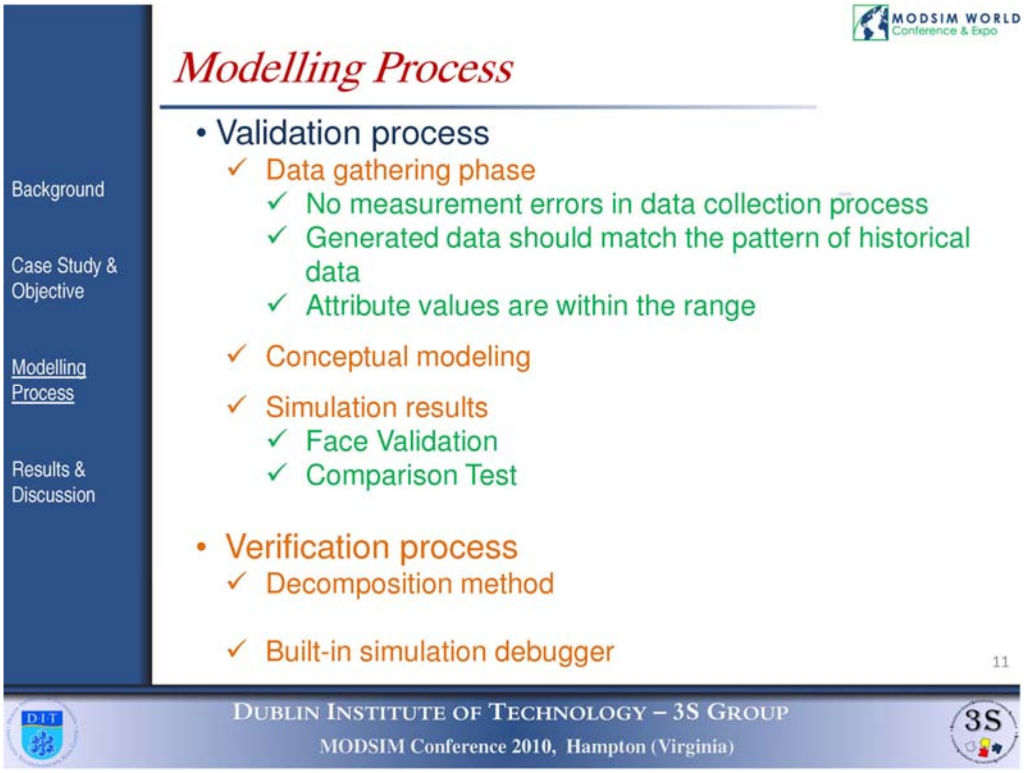
Modelling Process

Results & Discussion

10

DUBLIN INSTITUTE OF TECHNOLOGY – 3S GROUP

MODSIM Conference 2010, Hampton (Virginia)



Modelling Process

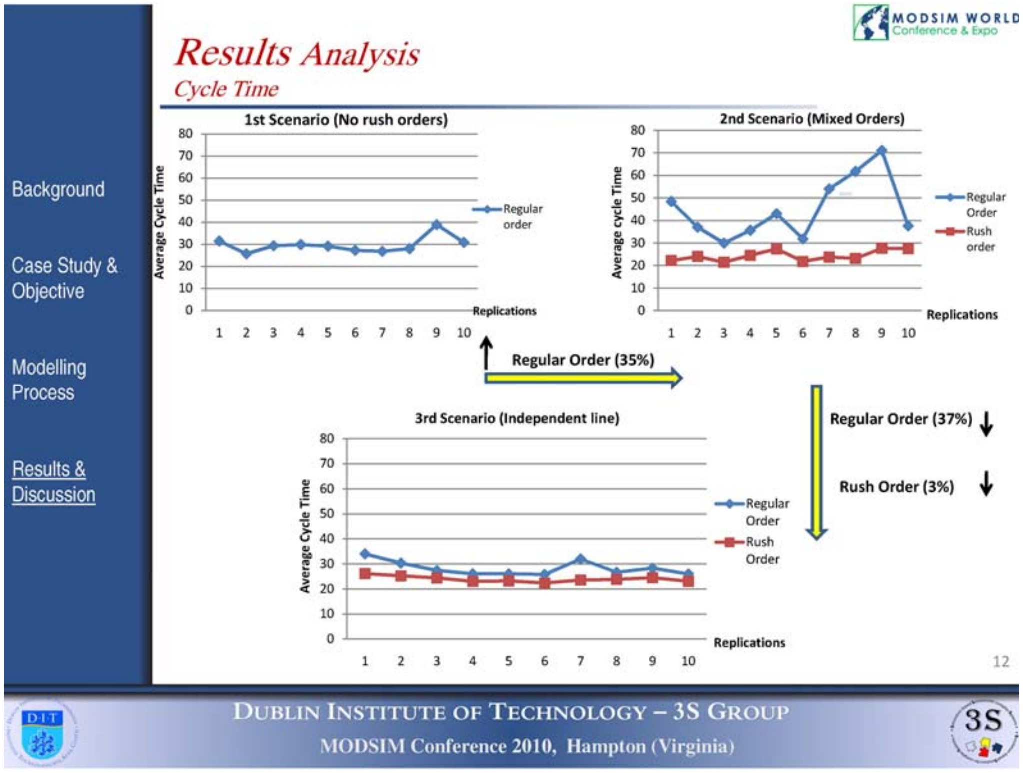
Background

Case Study & Objective

Modelling Process

Results & Discussion

- Validation process
  - ✓ Data gathering phase
    - ✓ No measurement errors in data collection process
    - ✓ Generated data should match the pattern of historical data
    - ✓ Attribute values are within the range
  - ✓ Conceptual modeling
  - ✓ Simulation results
    - ✓ Face Validation
    - ✓ Comparison Test
- Verification process
  - ✓ Decomposition method
  - ✓ Built-in simulation debugger



Results Analysis

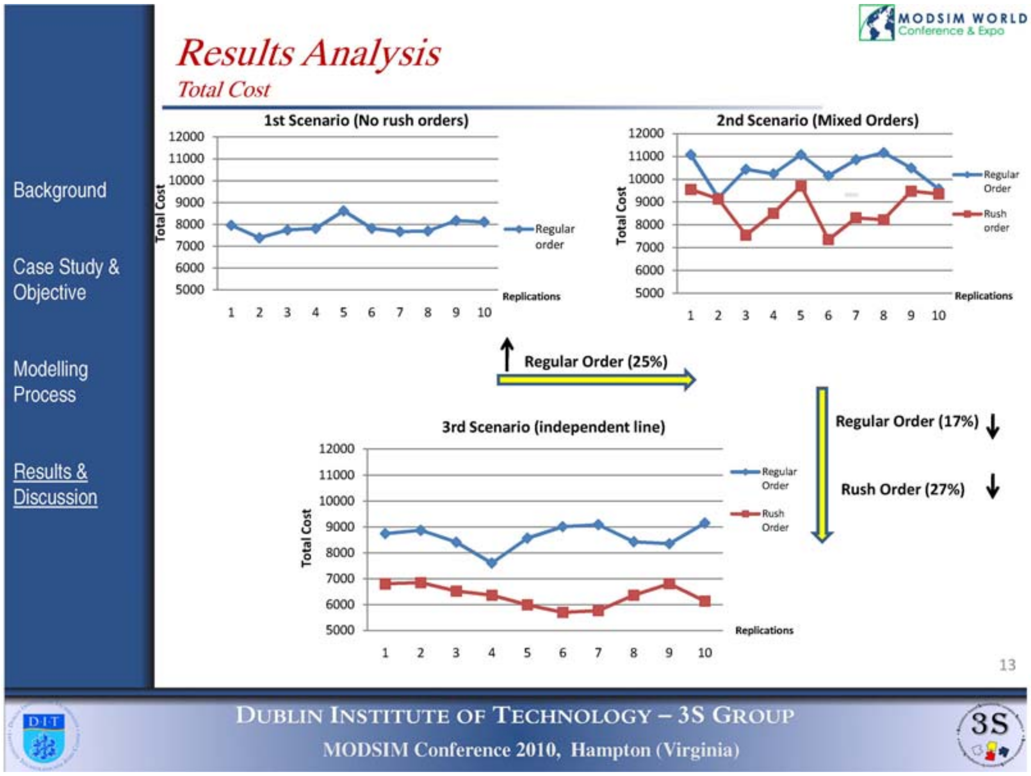
Cycle Time

Background

Case Study & Objective

Modelling Process

Results & Discussion



MODSIM WORLD Conference & Expo

Results Analysis

Total Cost

Background

Case Study & Objective

Modelling Process

Results & Discussion

↑ Regular Order (25%)

Regular Order (17%) ↓

Rush Order (27%) ↓

13

DUBLIN INSTITUTE OF TECHNOLOGY – 3S GROUP

MODSIM Conference 2010, Hampton (Virginia)

MODSIM WORLD Conference & Expo

Background

Case Study & Objective

Modelling Process

Results & Discussion

Thanks

&

Any Questions ?

14

DUBLIN INSTITUTE OF TECHNOLOGY – 3S GROUP

MODSIM Conference 2010, Hampton (Virginia)

2.7 Meta-RaPS Algorithm for the Aerial Refueling Scheduling Problem

Meta-RaPS Algorithm for the Aerial Refueling Scheduling Problem

Sezgin Kaplan; Arif Arin; Ghaith Rabadi
Old Dominion University
Engineering Management and Systems Engineering
skaplan@odu.edu; aarin@odu.edu; grabadi@odu.edu

Abstract. The Aerial Refueling Scheduling Problem (ARSP) can be defined as determining the refueling completion times for each fighter aircraft (job) on multiple tankers (machines). ARSP assumes that jobs have different release times and due dates. The total weighted tardiness is used to evaluate schedule's quality. Therefore, ARSP can be modeled as a parallel machine scheduling with release times and due dates to minimize the total weighted tardiness. Since ARSP is NP-hard, it will be more appropriate to develop approximate or heuristic algorithm to obtain solutions in reasonable computation times. In this paper, Meta-Raps-ATC algorithm is implemented to create high quality solutions. Meta-RaPS (Meta-heuristic for Randomized Priority Search) is a recent and promising metaheuristic that is applied by introducing randomness to a construction heuristic. The Apparent Tardiness Rule (ATC), which is a good rule for scheduling problems with tardiness objective, is used to construct initial solutions which are improved by an exchanging operation. Results are presented for generated instances.

1.0 INTRODUCTION

Resources commonly occur in parallel and many real life problems can be modeled as parallel machine scheduling problems. A parallel machine scheduling problem involves both resource allocation and sequencing. It allocates jobs to each machine and determines the sequence of allocated jobs on each machine. Aerial refueling (AR) is the process of transferring fuel from a tanker aircraft to another receiver aircraft during flight. Aerial refueling scheduling problem (ARSP) aims to determine the starting and completion times of refueling process of each receivers on the tankers. ARSP can be modeled as an identical parallel machine scheduling problem with release times and due dates. It represents a system with *m* identical machines in parallel and *n* jobs where job *j* arrives (becomes available) at ready time r_j and should be complete and leave by the due date d_j. The objective is to find the schedule minimizing total weighted tardiness (TWT) as a performance measure to maintain the quality of service with due dates.

Since ARSP is NP-hard from complexity point of view, it is required to develop effective solution approaches with reasonable computation times. In this study, a fairly new metaheuristic, Meta-RaPS, will be applied to solve the ARSP. Meta-RaPS stands for "Meta-heuristic for Randomized Priority Search", and is one of the randomized search metaheuristics. DePuy et al. [1] expresses the advantages of the Meta-RaPS over other metaheuristics. According to them, run times for Meta-RaPS is not significantly affected by the size of the problem, it is easy to understand and to implement, and can generate a feasible solution at every iteration. It requires a simple dispatching rule to randomize and escape local optima.

Dispatching (or Priority) Rules are the most common heuristics for scheduling problems due to their easy implementation and low computational requirements. The Apparent Tardiness Cost (ATC) heuristic is a good composite dispatching rule for the parallel machine total weighted tardiness problem and is used with MetaRaPS in this paper.

The rest of this paper is organized as follows. In Section 2, the related research is summarized. The Meta-RaPS metaheuristic is explained in Section 3 and the ATC rule in Section 4. A computational study is described in Section 5 by giving an example of the construction phase calculations and a comparison of the TWT values obtained by Meta-RaPS-ATC algorithm with the values obtained by ATC alone. Finally results are concluded in Section 6.

2.0 RELATED WORK

Some researchers addressed scheduling identical parallel machines with ready times to minimize total weighted tardiness problem. Mönch et al. [2] attempted to minimize total weighted tardiness on parallel batch machines with incompatible job families and unequal ready times. They proposed two different decomposition approaches. Dispatching and scheduling rules were used for the batching phase and the sequencing phase of the two approaches. Reichelt et al. [3] were interested in minimizing total weighted tardiness and makespan at the same time. In order to determine a pareto efficient solution for the scheduling of jobs with incompatible families on parallel batch machines problem, they suggested a hybrid multi objective genetic algorithm. Pfund et al. [4] addressed scheduling jobs with ready times on identical parallel machines with sequence dependent setups by minimizing the total weighted tardiness. Their approach was an extension of the Apparent Tardiness Cost with Setups (ATCS) approach by Lee and Pinedo [5] to allow non-ready jobs to be scheduled. Gharehgozli et al. [6] presented a new mixed-integer goal programming (MIGP) model for a parallel machine scheduling problem with sequence-dependent setup times and release dates. Fuzzy processing times and two fuzzy objectives were considered in the model to minimize the total weighted flow time and the total weighted tardiness simultaneously.

There are also a few Meta-RaPS applications on scheduling problems. Hepdogan et al. [7] investigated Meta-RaPS approach to the single machine early/tardy scheduling problem with common due date and sequence-dependent setup times. The objective of their problem was to minimize the total amount of earliness and tardiness of jobs that are assigned to a single machine. Rabadi et al. [8] introduced Meta-RaPS approach to the non-preemptive unrelated parallel machine scheduling problem with the objective of minimizing the makespan. In their problem, machine-dependent and job sequence-dependent setup times were considered when all jobs are available at time zero, and all times are deterministic.

3.0 THE META-RAPS ALGORITHM

Moraga et al. [9] defines Meta-RaPS as "generic, high level search procedures that introduce randomness to a construction heuristic as a device to avoid getting trapped at a local optimal solution". Meta-RaPS combines the mechanisms of priority rules, randomness, and sampling.

A Meta-RaPS algorithm uses four parameters: the number of iterations (*I*), the priority percentage (p%), the restriction percentage (*r*%), and the improvement percentage (*i*%). Meta-RaPS does not select the component or activity with the best priority value every time, nor the incremental cost. However, the algorithm may accept one with a good priority value, not necessarily the best, based on a randomized approach. The parameter *p*% is employed to decide the percentage of time, the component, or activity with the best priority value will be added to the current partial solution, and 100%-*p*% of time the component or activity with the good priority value is randomly selected from a candidate list (CL) containing "good" components or activities. The CL of components or activities with good priority values is created by including ones whose priority values are within *r*% of the best priority value.

Meta-RaPS is a two-phase metaheuristic: a constructive phase to create feasible solutions and an improvement phase to improve them. In the constructive phase, a solution is built by repeatedly adding feasible components or activities to the current solution in order based on their priority rules until the stopping criterion is satisfied. Generally, solutions obtained by implementing only constructive algorithms can reach mostly local optima. To avoid local optima, Meta-RaPS employs randomness in the constructive phase so that solutions other than the best solution can be selected.

The improvement phase is performed if the feasible solutions generated in the construction phase are within *i*% of the best unimproved solution value from the preceding iterations [9].

4.0 ATC RULE

The Apparent Tardiness Cost (ATC) will be applied to the parallel machine total weighted tardiness problem as a composite dispatching rule. Pinedo [10] defines a composite dispatching rule as "a ranking expression that combines a number of elementary dispatching rules". An elementary rule is a function of constant or time dependent properties of the jobs and/or the machines, i.e. processing times, due dates for jobs; speed, number of jobs waiting for processing for machines, etc. The ATC combines the elementary Weighted Shortest Processing Time first (WSPT) dispatching rule and the Minimum Slack first (MS) rule. According to the WSPT rule the jobs are ordered in decreasing order of w_j/p_j; and the MS rule selects at time *t*, when a machine is freed, among the remaining jobs the job with the minimum slack where the slack can be defined as $max\ (d_j - p_j - t,\ 0)$. Every time the machine becomes free, the ATC calculates a ranking index for each remaining job. The job with the highest ranking index defined in equation 1 is then selected to be processed next:

$$I_j(t) = \frac{w_j}{p_j} \exp\left(-\frac{\max(d_j - p_j - t, 0)}{K\bar{p}} \right) \qquad (1)$$

where $\bar{p}$ is the average processing time of the remaining jobs, and *K* is the scaling parameter, called look-ahead parameter. If *K* is very large the ATC rule behaves similar to the WSPT rule, and if *K* is very small the rule behaves similar to the MS rule. The WSPT rule is optimal when all jobs are tardy, while the MS rule is optimal when all due dates are sufficiently loose and spread out.

The effectiveness of the ATC heuristic depends on the value of the look-ahead parameter *K*. Previous studies have usually recommended a fixed value of *K* between 0.5 and 2.0 [11].

5.0 COMPUTATIONAL STUDY

5.1 Parameter Setting

The accepted values of the parameters to be employed in metaheuristics have a significant impact on both the solution process and solution quality. Particularly, in terms of the interactions, Design of Experiments (DOE) methods are promising approaches and can be employed to tune the parameters more effectively. In this study, we applied 3-level (3^k) full factorial design to tune the parameters of Meta-RaPS. After completing regression analysis with R^2 = 0.95, the values found for the parameters are presented in Table 1.

Table 1. Meta-RaPS Parameter Setting

Parameter	Value
Number of iterations (I)	10000
Priority percentage (p%)	25%
Restriction percentage (r%)	60%
Improvement percentage (i%)	70%

5.2 Meta-RaPS-ATC Algorithm

To present the effectiveness of Meta-RaPS-ATC algorithm, problems were solved both by using ATC rule and Meta-RaPS-ATC approach with the tuned parameters. In the ATC, the jobs are selected by calculating their ATC index, and the one with the highest index is always selected. However, in Meta-RaPS-ATC algorithm, the ATC index for each job is calculated, and the selection is made based on Meta-RaPS principles. If the random number (RN) is smaller or equal to the priority percentage, the job with the highest ATC index is selected. If not, a lower limit is calculated by multiplying the highest index by the restriction percentage. Jobs whose ATC indices are higher than the lower limit are added to the CL, and the next job is selected from this CL randomly. After all jobs are assigned to machines, the construction phase of Meta-RaPS is completed. For both algorithms, the ATC indices are updated after the selection of each job.

In Meta-RaPS, only the constructed solutions with promising, or good enough, values are improved. To determine this level, summation of the lowest (best) solution with the multiplication of the difference between the highest (worst) solution value and the best solution value obtained until current iteration by the improvement percentage is used. If the current solution value is higher than this level, the improvement phase is performed by swapping two arbitrarily selected jobs in the constructed schedule and comparing with the best and the worst solution values in memory. After swapping operation, jobs are scheduled by taking into account the release time of the swapped job and the completion time of the predecessor job.

5.3 Results for ARSP

To simulate ARSP, we randomly generated 10 instances with release times, processing times, weights and due dates for m = 3 machines and n = 12 jobs so that an optimal solution can be obtained in a reasonable time. One of these instances whose data is given in Table 2. is used as an example to explain Meta-RaPS.

Table 2. Data for Example Problem

Job	Release Time	Processing Time	Weight	Due Date
1	19	30	4	79
2	1	28	7	55
3	24	23	8	70
4	1	28	8	55
5	3	19	6	41
6	10	45	5	100
7	2	45	4	92
8	5	43	3	91
9	11	28	1	67
10	0	23	8	46
11	13	29	1	71
12	15	32	2	79

The construction phase of Meta-RaPS algorithm only for one iteration is shown in Table 3. In every step of this phase, jobs are assigned to the machines with the earliest availability. The total weighted tardiness of this solution is 250.

A Mixed Integer Linear Programming (MILP) model was developed to find optimal solutions. Optimization Programming Language (OPL) Studio 6.3 was used to implement this model and CPLEX 12.1 to solve it.

Table 3. Construction Phase of Meta-RaPS ATC Algorithm

	Machine			Max.		Lower						
Step	1	2	3	ATC	Job	Limit	CL	RN	RN > p	Assignment	Job	Machine
1	0	0	0	0.187	10	0.112	4, 7	0,21	NO	max	10	1
2	23	0	0	0,177	5	0,106	2,4	0,85	YES	CL	2	2
3	23	28	0	0,178	5	0,107	4	0,77	YES	CL	4	3
4	23	28	28	0,316	5	0,189	empty	-	-	max	5	1
5	42	28	28	0,219	3	0,132	empty	-	-	max	3	2
6	42	51	28	0,082	1	0,049	6,7	0,51	YES	CL	7	3
7	42	51	73	0,113	1	0,068	6,7	0,16	NO	max	1	1
8	72	51	73	0,101	6	0,061	8	0,23	NO	max	6	2
9	72	96	73	0,021	8	0,012	12	0,69	YES	CL	12	1
10	104	96	73	0,073	8	0,044	empty	-	-	max	8	3
11	104	96	116	0,036	9	0,021	11	0,22	NO	max	9	2
12	104	124	116	-	-	-	-	-	-	-	11	1

Table 4. Comparison of ATC and Meta-RaPS ATC Algorithm

Instance	ATC	Meta-RaPS	Optimal	ATC Deviation from Optimal	Meta-RaPS-ATC Deviation from Optimal
1	707.5	495.3	471.5	0.50	0.05
2	248.0	218.6	216.0	0.15	0.01
3	790.0	770.0	680.0	0.16	0.13
4	270.9	265.8	238.0	0.14	0.12
5	478.1	364.0	363.0	0.32	0.00
6	484.0	394.5	372.0	0.30	0.06
7	276.0	278.6	276.0	0.00	0.01
8	211.5	116.6	100.5	1.10	0.16
9	315.7	296.0	287.0	0.10	0.03
10	526.2	364.6	348.0	0.51	0.05
			Average	**0.32**	**0.06**

The results for 10 instances are summarized in Table 4. While the average deviation of the ATC results from the optimal solutions was 0.32, and the average deviation of the Meta-RaPS-ATC algorithm solutions was 0.06. Based on the findings for the ARSP instances, using Meta-RaPS-ATC approach gives better results than using ATC rule.

6.0 CONCLUSIONS

ARSP is a real world problem that requires high quality solutions in an acceptable time frame. As the dimensions of the problem get larger, the solution process of mathematical modeling loses its effectiveness. Using only composite dispatching rules, such as the ATC rule, may not give the best solutions for most applications. However, metaheuristics can offer high quality solutions, and Meta-RaPS seems to be a promising metaheuristic with its simplicity and effectiveness to find high quality solutions for ARSP, and for scheduling problems in general.

More computation and analysis are needed for better performance comparisons in instances with large number of jobs. In future research, more constraints such as machine compatibility, sequence dependent setup times and deadlines may be included in the model.

7.0 REFERENCES

1. DePuy G. W., Whitehouse G. E. and Moraga R. J., Meta-RaPS: A Simple And Efficient Approach For Solving Combinatorial Problems, 29th International Conference on Computers and Industrial Engineering, November 1-3, Montreal, Canada, 644-649, 2001.

2. Mönch, L., Balasubramanian, H., Fowler, J. W. and Pfund, M.E., Heuristic scheduling of jobs on parallel batch machines with incompatible job families and unequal ready times, Computers & Operations Research, 32, 2731–2750, 2005.

3. Reichelt, D. , Mönch, L., Gottlieb, J. and Raidl, G.R., Multiobjective Scheduling of Jobs with Incompatible Families on Parallel Batch Machines, EvoCOP 2006, LNCS 3906, Berlin Springer-Verlag Heidelberg, 209–221, 2006.

4. Pfund, M., Fowler, J. W., Gadkari, A. and Chen, Y., Scheduling jobs on parallel machines with setup times and ready times, Computers and Industrial Engineering, 54, 764–782, 2008.

5. Lee, Y.H. and Pinedo, M., Theory and methodology: Scheduling jobs on parallel machines with sequence-dependent setup times, European Journal of Operational Research 100, 464-474, 1997.

6. Gharehgozli, A.H., Tavakkoli, R. and Zaerpour, N., A fuzzy-mixed-integer goal programming model for a parallel-machine scheduling problem with sequence-dependent

setup times and release dates, Robotics and Computer-Integrated Manufacturing, 25, 853–859, 2009.

7. Hepdogan S., Moraga R., DePuy G.W., and Whitehouse G.E, A Meta-RaPS for the early/tardy single machine scheduling problem, International Journal of Production Research, Vol. 47, No. 7, 1717–1732, 2009.

8. Rabadi G., Moraga R.J. and Al-Salem A., Heuristics for the unrelated parallel machine scheduling problem with setup times, Journal of Intelligent Manufacturing, 17, 85–97, 2006.

9. Moraga R. J., DePuy G. W. and Whitehouse G. E., Metaheuristics: A Solution Methodology for Optimization Problems, Handbook of Industrial and Systems Engineering, CRC Press, FL, 2006.

10. Pinedo M. L., Scheduling Theory, Algorithms, and Systems, Third Edition, Springer, NY, 2008.

11. Holsenback J. E., Russell R. M., Markland R. E. and Philipoom P. R., An improved heuristic for the single-machine, weighted-tardiness problem, Omega 27, 485–495, 1999.

MODSIM WORLD Conference & Expo

October 13–15, 2010
Hampton, Virginia

Meta-RaPS Algorithm for the Aerial Refueling Scheduling Problem

Sezgin KAPLAN, Ph.D. Student
Old Dominion University

October 13, 2010

MODSIM WORLD Conference & Expo

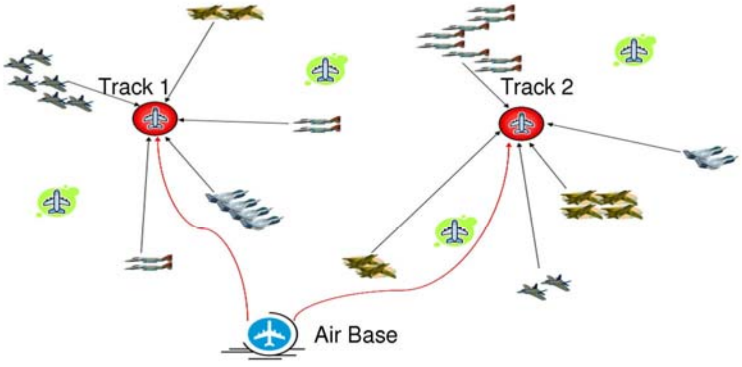
Outline

- Aerial Refueling Scheduling Problem (ARSP)
- ARSP Solution Method
- Related Works
- Meta-RaPS Algorithm
- Apparent Tardiness Cost (ATC) Rule
- Computational Study
- Conclusion
- References

ARSP

Aerial Refueling: The process of transferring fuel from one aircraft (a tanker) to another (a fighter as a receiver) during flight.

ARSP



- ARSP: Determining the refueling completion times for each fighter aircraft (job) on one of multiple tankers (machines).

MODSIM WORLD
Conference & Expo

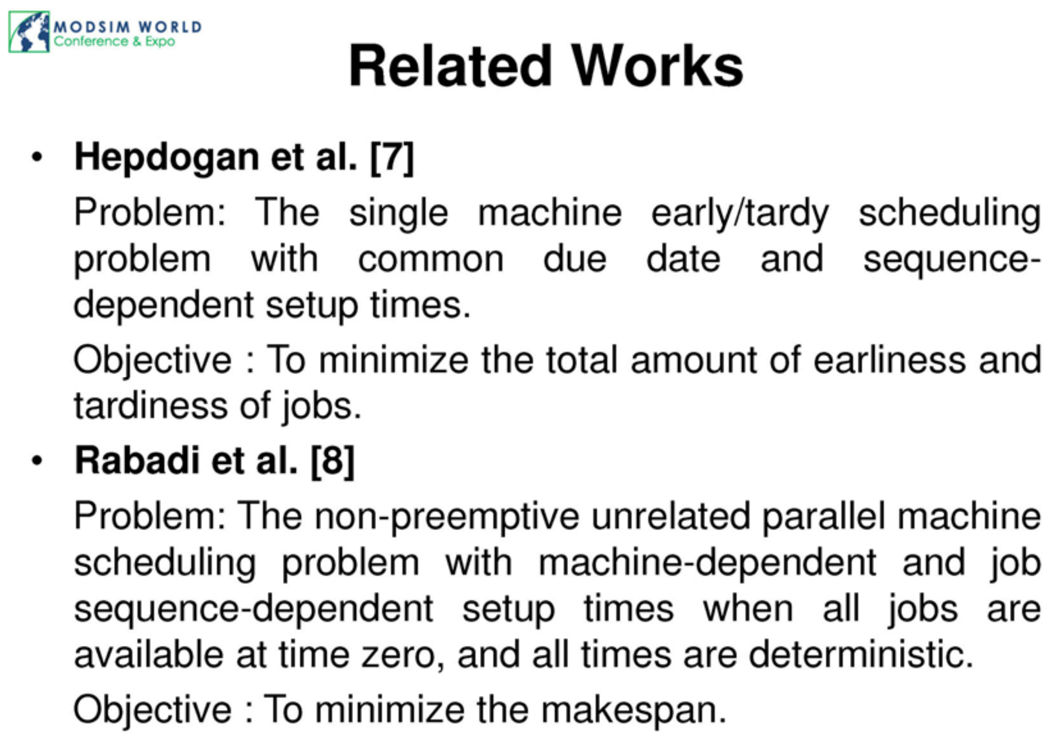
ARSP

- **Model :** An identical parallel machine scheduling problem with release times and due dates.
- A system with *m* identical machines in parallel and *n* jobs where job *j* becomes available at ready time r_j and should be complete and leave by the due date d_j.
- **Objective :** To minimize total weighted tardiness (TWT) as a performance measure.

MODSIM WORLD
Conference & Expo

Solution Method

- MetaRaPS- ATC algortihm will be applied to solve the ARSP.
- "Meta-heuristic for Randomized Priority Search" (Meta-RaPS) is;
  - One of the randomized search metaheuristics.
  - It requires a simple dispatching rule to randomize and escape local optima [1].
- The Apparent Tardiness Cost (ATC) heuristic is used as a dispatching rule with MetaRaPS.



Related Works

- **Hepdogan et al. [7]**

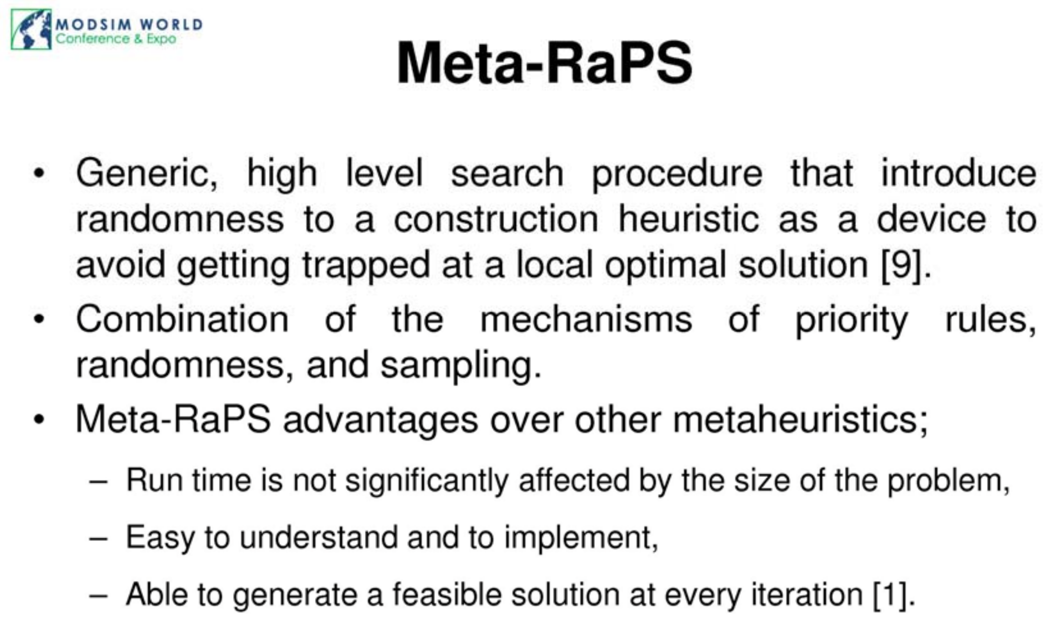
  Problem: The single machine early/tardy scheduling problem with common due date and sequence-dependent setup times.

  Objective : To minimize the total amount of earliness and tardiness of jobs.

- **Rabadi et al. [8]**

  Problem: The non-preemptive unrelated parallel machine scheduling problem with machine-dependent and job sequence-dependent setup times when all jobs are available at time zero, and all times are deterministic.

  Objective : To minimize the makespan.



Meta-RaPS

- Generic, high level search procedure that introduce randomness to a construction heuristic as a device to avoid getting trapped at a local optimal solution [9].
- Combination of the mechanisms of priority rules, randomness, and sampling.
- Meta-RaPS advantages over other metaheuristics;
  - Run time is not significantly affected by the size of the problem,
  - Easy to understand and to implement,
  - Able to generate a feasible solution at every iteration [1].

Meta-RaPS

Two-phase in each iteration:

1. **Constructive phase** to create feasible solutions.
   - A solution is built by repeatedly adding feasible activities to the current solution in order based on their priority rules.
   - To avoid local optima, randomness is employed so that solutions other than the best solution can be selected.
2. **Improvement phase** to improve the feasible solution.
   - Only the constructed solutions with promising, or good enough, values are improved.
   - A neighborhood search is performed.

Meta-RaPS

Meta-RaPS algorithm uses four parameters:

1. **Priority percentage** (p%): Decision to schedule the best or randomly one from candidate list.
2. **Restriction percentage** (*r*%) : Determining the CL by calculating a lower limit for good priority values. The lower limit is *r*% of the best priority value.
3. **Improvement percentage** (*i*%) : Decision to improve or not improve the constructed solution.
4. **Number of iterations** (*I*) : Stopping criterion

Meta-RaPS

The pseudocode for one iteration of the basic Meta-RaPS procedure is given as follows:

Construction Phase

```
Do until feasible solution generated
    Find priority value for each feasible activity
    Find best ATC priority value
    P = RND(1,100)
    If P ≤ %priority Then
        Add activity with best priority value to solution
    Else
        Form “available” list of all feasible activities whose priority values
         are within %restriction of best priority value
        Randomly choose activity from available list and add to solution
    End If
End Until
```

Improvement Phase

```
If an iteration’s solution value is within %improvement of the best solution
value found so far, Then
    A neighborhood search is performed
End If
Calculate and Print solution
```

ATC Rule

- A good composite dispatching rule for the total weighted tardiness problem [4].
- A dynamic algorithm that after each job completion, the remaining jobs are analyzed, priorities derived according to improved formula, and the highest priority job selected.
- The priority index :

$$\pi_j(t) = \frac{w_j}{p_j}\exp\left[-\frac{\max\left[d_j - p_j - t, 0\right]}{k\overline{p}}\right]$$

  - Urgent job :

    If $C_j \geq d_j$, then $\pi_j = w_j/p_j$

w_j : weight of the remaining job j

p_j : processing time of the job j

d_j : due date of the job j

t : Decision time point

$\overline{p}$: The average processing time of the remaining jobs,

k : ‘look-ahead‘ parameter.

$S_j^+(t)$, the slack factor : $max\ [d_j - p_j - t, 0]$.

Parameter Setting

- The accepted values of the parameters to be employed in metaheuristics have a significant impact on both the solution process and solution quality.
- Design of Experiments (DOE) methods are used to tune the parameters.

Table 3. Meta-RaPS Parameters

Parameter	Value
Number of iterations (I)	10000
Priority percentage (p%)	25%
Restriction percentage (r%)	60%
Improvement percentage (i%)	70%

An Example

Table 1. Data for Example Problem (*m* = 3 machines and *n* = 12 jobs)

Job	Release Time	Processing Time	Weight	Due Date
1	19	30	4	79
2	1	28	7	55
3	24	23	8	70
4	1	28	8	55
5	3	19	6	41
6	10	45	5	100
7	2	45	4	92
8	5	43	3	91
9	11	28	1	67
10	0	23	8	46
11	13	29	1	71
12	15	32	2	79

An Example

Table 2. Construction Phase of Meta-RaPS ATC Algorithm

	Machine			Max.		Lower						
						Limit	CL	RN	RN > p	Assignment	Job	Machine
Step	1	2	3	ATC	Job	r= 0.6			p= 0.25			
1	0	0	0	0.187	10	0.112	4,7,10	0.21	NO	max	10	1
2	23	0	0	0,177	5	0,106	2,4,5	0.85	YES	CL	2	2
3	23	28	0	0,178	5	0,107	4,5	0.77	YES	CL	4	3
4	23	28	28	0,316	5	0,189	5	-	-	max	5	1
5	42	28	28	0,219	3	0,132	3	-	-	max	3	2
6	42	51	28	0,082	1	0,049	1,6,7	0.51	YES	CL	7	3
7	42	51	73	0,113	1	0,068	1,6,7	0.16	NO	max	1	1
8	72	51	73	0,101	6	0,061	6,8	0.23	NO	max	6	2
9	72	96	73	0,021	8	0,012	8,12	0.69	YES	CL	12	1
10	104	96	73	0,073	8	0,044	8	-	-	max	8	3
11	104	96	116	0,036	9	0,021	9,11	0.22	NO	max	9	2
12	104	124	116	-	-	-	-	-	-	-	11	1

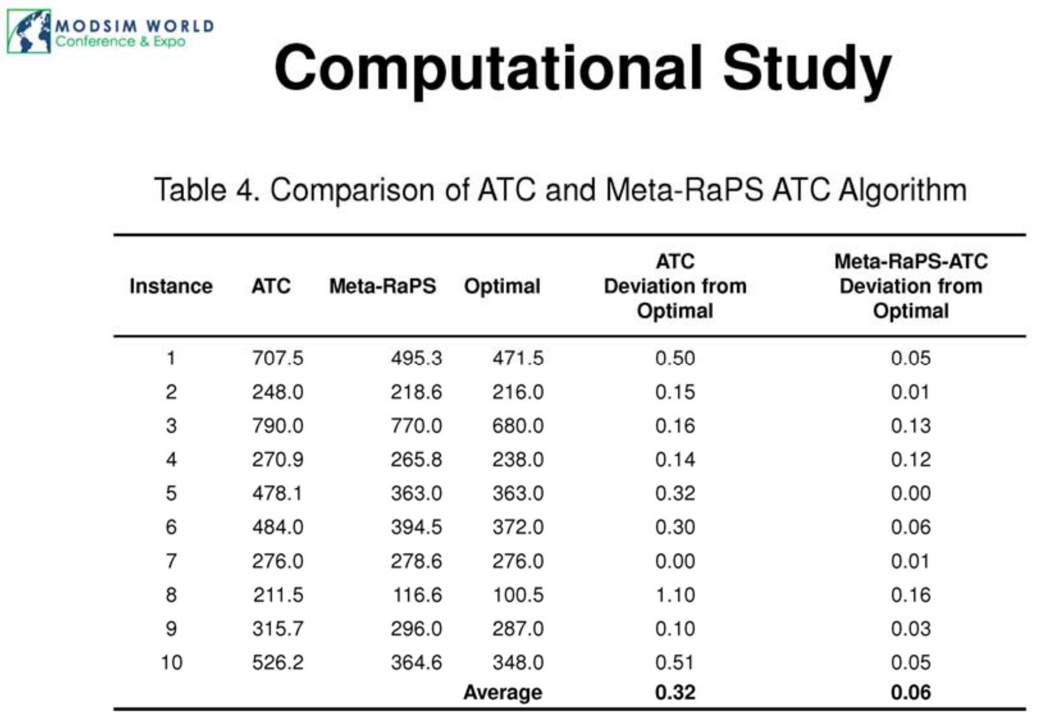
Machine 1: 10-5-1-12-11 **Machine 2**: 2-3-6-9 **Machine 3** : 4-7-8 **TWT** : 250

Assume Best : 200, Worst : 400, *%i* : 0.7
Decision level: 200 + (400-200)*0.7 = 340 > TWT Improve

Improvement : Swapping two arbitrarily selected jobs in the constructed schedule by taking into account the release times.

Computational Study

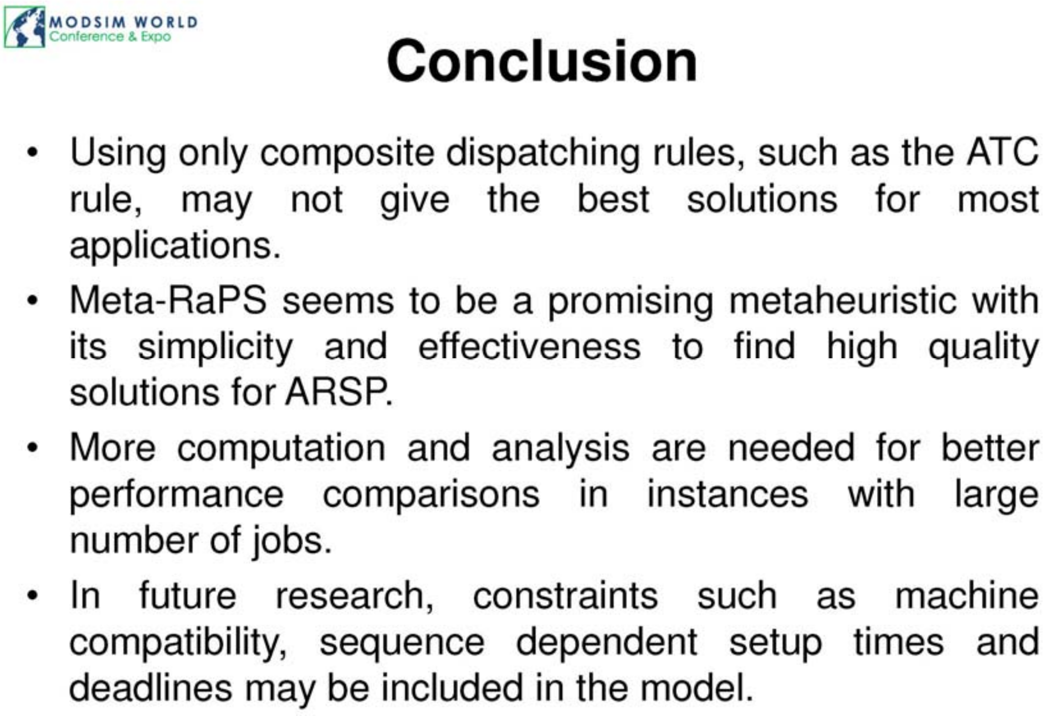
- To present the effectiveness of Meta-RaPS algorithm, problems were solved both by using ATC rule and Meta-RaPS algorithm with the tuned parameters.
- To simulate ARSP, 10 instances with release times, processing times, weights and due dates were randomly generated.
- A Mixed Integer Linear Programming (MILP) model was developed to find optimal solutions. Optimization Programming Language (OPL) Studio 6.3 was used to implement this model and CPLEX 12.1 to solve it.



Computational Study

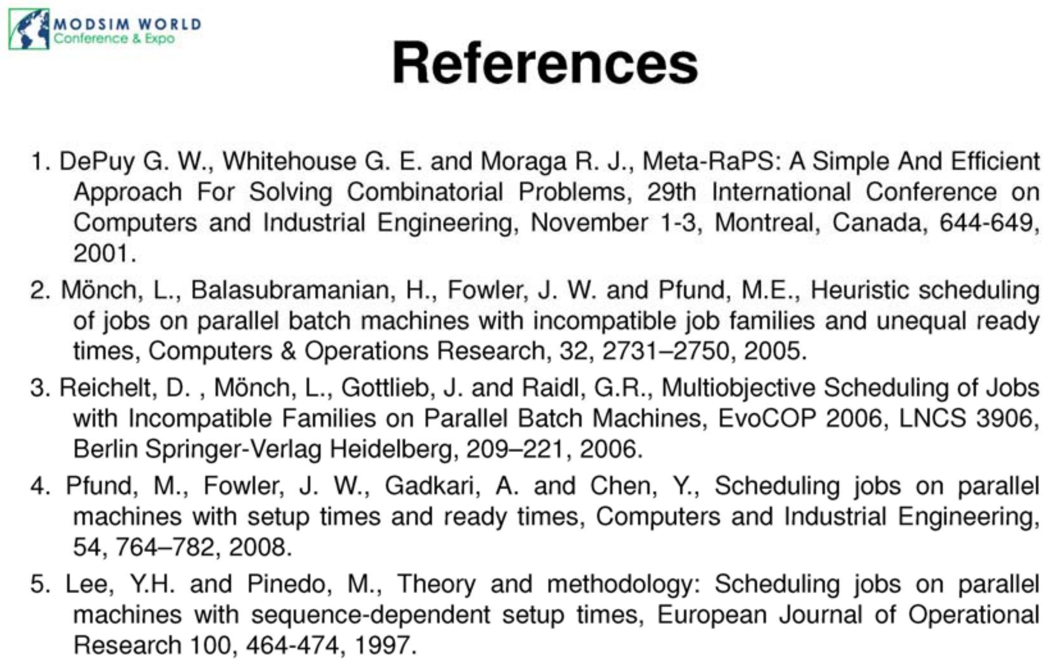
Table 4. Comparison of ATC and Meta-RaPS ATC Algorithm

Instance	ATC	Meta-RaPS	Optimal	ATC Deviation from Optimal	Meta-RaPS-ATC Deviation from Optimal
1	707.5	495.3	471.5	0.50	0.05
2	248.0	218.6	216.0	0.15	0.01
3	790.0	770.0	680.0	0.16	0.13
4	270.9	265.8	238.0	0.14	0.12
5	478.1	363.0	363.0	0.32	0.00
6	484.0	394.5	372.0	0.30	0.06
7	276.0	278.6	276.0	0.00	0.01
8	211.5	116.6	100.5	1.10	0.16
9	315.7	296.0	287.0	0.10	0.03
10	526.2	364.6	348.0	0.51	0.05
			Average	**0.32**	**0.06**



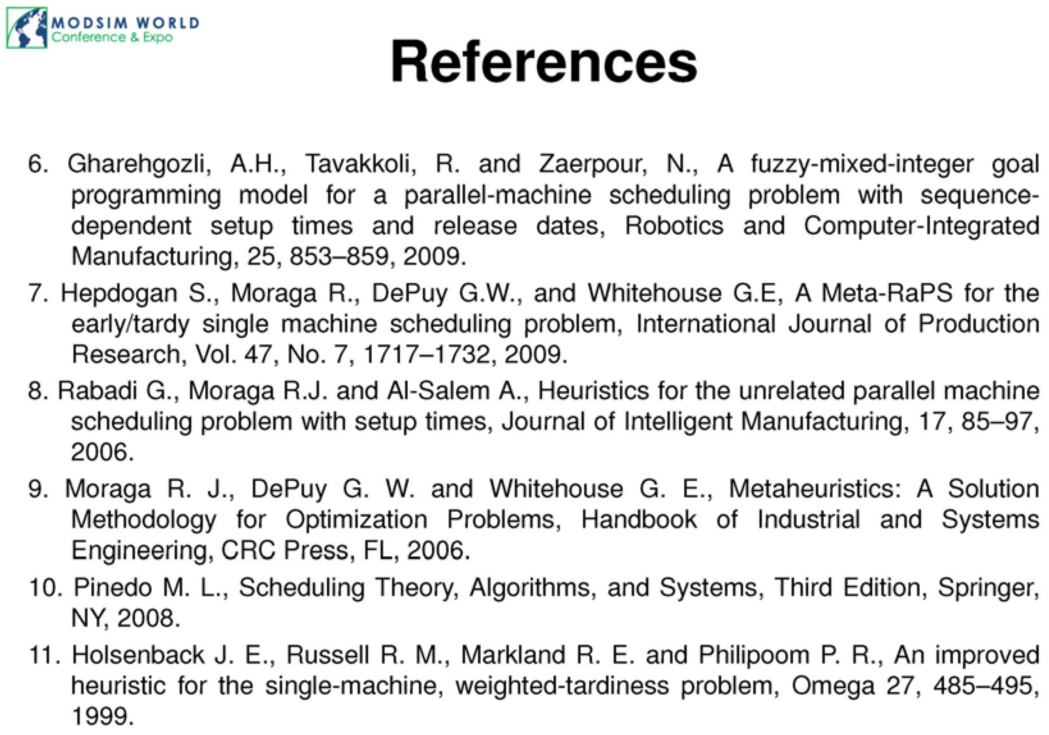
Conclusion

- Using only composite dispatching rules, such as the ATC rule, may not give the best solutions for most applications.
- Meta-RaPS seems to be a promising metaheuristic with its simplicity and effectiveness to find high quality solutions for ARSP.
- More computation and analysis are needed for better performance comparisons in instances with large number of jobs.
- In future research, constraints such as machine compatibility, sequence dependent setup times and deadlines may be included in the model.



References

1. DePuy G. W., Whitehouse G. E. and Moraga R. J., Meta-RaPS: A Simple And Efficient Approach For Solving Combinatorial Problems, 29th International Conference on Computers and Industrial Engineering, November 1-3, Montreal, Canada, 644-649, 2001.
2. Mönch, L., Balasubramanian, H., Fowler, J. W. and Pfund, M.E., Heuristic scheduling of jobs on parallel batch machines with incompatible job families and unequal ready times, Computers & Operations Research, 32, 2731–2750, 2005.
3. Reichelt, D. , Mönch, L., Gottlieb, J. and Raidl, G.R., Multiobjective Scheduling of Jobs with Incompatible Families on Parallel Batch Machines, EvoCOP 2006, LNCS 3906, Berlin Springer-Verlag Heidelberg, 209–221, 2006.
4. Pfund, M., Fowler, J. W., Gadkari, A. and Chen, Y., Scheduling jobs on parallel machines with setup times and ready times, Computers and Industrial Engineering, 54, 764–782, 2008.
5. Lee, Y.H. and Pinedo, M., Theory and methodology: Scheduling jobs on parallel machines with sequence-dependent setup times, European Journal of Operational Research 100, 464-474, 1997.



References

6. Gharehgozli, A.H., Tavakkoli, R. and Zaerpour, N., A fuzzy-mixed-integer goal programming model for a parallel-machine scheduling problem with sequence-dependent setup times and release dates, Robotics and Computer-Integrated Manufacturing, 25, 853–859, 2009.
7. Hepdogan S., Moraga R., DePuy G.W., and Whitehouse G.E, A Meta-RaPS for the early/tardy single machine scheduling problem, International Journal of Production Research, Vol. 47, No. 7, 1717–1732, 2009.
8. Rabadi G., Moraga R.J. and Al-Salem A., Heuristics for the unrelated parallel machine scheduling problem with setup times, Journal of Intelligent Manufacturing, 17, 85–97, 2006.
9. Moraga R. J., DePuy G. W. and Whitehouse G. E., Metaheuristics: A Solution Methodology for Optimization Problems, Handbook of Industrial and Systems Engineering, CRC Press, FL, 2006.
10. Pinedo M. L., Scheduling Theory, Algorithms, and Systems, Third Edition, Springer, NY, 2008.
11. Holsenback J. E., Russell R. M., Markland R. E. and Philipoom P. R., An improved heuristic for the single-machine, weighted-tardiness problem, Omega 27, 485–495, 1999.

MODSIM WORLD
Conference & Expo

Questions/Comments

2.8 Aerial Refueling Process Rescheduling Under Job Related Disruptions

Aerial Refueling Process Rescheduling Under Job Related Disruptions

Sezgin Kaplan & Ghaith Rabadi
Old Dominion University
skaplan@odu.edu grabadi@odu.edu

Abstract. The Aerial Refueling Scheduling Problem (ARSP) can be defined as determining the refueling completion times for each fighter aircraft (job) on the multiple tankers (machines) to minimize the total weighted tardiness. ARSP assumes that the jobs have different release times and due dates. The ARSP is dynamic environment and unexpected events may occur. In this paper, rescheduling in the aerial refueling process with a finite set of jobs will be studied to deal with job related disruptions such as the arrival of new jobs, the departure of an existing job, high deviations in the release times and changes in job priorities. In order to keep the stability and to avoid excessive computation, partial schedule repair algorithm is developed and its preliminary results are presented.

1.0 INTRODUCTION

Aerial refueling (AR) is the process of transferring fuel from one aircraft (the tanker) to another receiver aircraft during flight. The aerial refueling scheduling problem (ARSP) can be modeled as a parallel machine scheduling problem in which we need to determine which jobs have to be allocated to which machines and the sequence of the jobs allocated to each machine. ARSP aims to determine the starting and completion times of refueling process of each receivers on the tankers. It represents a system with *m* identical machines in parallel and *n* jobs where job *j* arrives (becomes available) at ready time r_j and should be complete and leave by the due date d_j.

There are some difficulties that make ARSP different from an ordinary parallel machine scheduling problem. One of these difficulties is sourced from a dynamic environment of the aerial refueling process where disruptions caused by dynamic and unexpected events require rescheduling to update the existing aerial refueling schedule.

In this paper, the parallel machine rescheduling problem with the multiple objectives of minimizing the total weighted tardiness and minimizing schedule instability will be addressed. It is assumed that schedules will be updated only as a result of job related disruptions such as the arrival of new jobs, the departure of an existing job, high deviations in the release times and changes in job priorities. The job related disruptions require a partial rescheduling procedure that aims to change only the affected part of the schedule in order to keep the stability and to avoid excessive computation. As a specific implementation of this procedure, a partial schedule repair algorithm for job arrival disruption has been developed.

The rest of this paper is organized as follows. In Section 2, the related research is summarized. In Section 3, a general rescheduling methodology for job related disruptions is explained. The partial schedule repair algorithm for job arrival disruption is introduced and a sample problem is given in Section 4. Comparison of the partial rescheduling with the regeneration (complete) rescheduling is described in Section 5, and finally results are concluded in Section 6.

2.0 RELATED WORKS

In the literature, rescheduling is required as a result of different disruptions and events such as new (rush) job arrivals, order cancellations, machine failure, changes in order priority, change in ready times, processing time delays, rework due to quality problems, material shortage, operator absenteeism, tool unavailability, due date changes, job order amount changes.

A few researches on the job related disruptions in parallel machine rescheduling can be listed to include Church and Uzsoy [1], Curry and Peters [2], Duenas and Petrovic [3]. Church and Uzsoy [1] introduced a combined periodic and event-driven approach. They developed worst-case error bounds for the periodic approach assuming that an optimal algorithm is used to reschedule the jobs available at each rescheduling point. Ref. [2] considered identical parallel machine scheduling problem with stepwise increasing tardiness cost objectives. Schedule nervousness increases when a scheduling procedure reassigns many planned operations to different machines or different start times. Their measure of schedule nervousness is the proportion of rescheduled jobs that change machine assignment. Ref. [3] presented a new predictive-reactive approach to identical parallel machine scheduling problem with material shortage and job arrival as an uncertain disruption. Their approach is based on generating a predictive schedule using dispatching rules to minimize the makespan. Two rescheduling methods namely left-shifting and building new schedules have been applied. The instability is measured as the starting time deviations between the predictive schedule and the reactive schedule.

Parallel machine rescheduling problems generally have multiple objectives: the objective of the original problem (e.g. minimization total weighted tardiness in our case) and the minimization of the difference between the new schedule (after rescheduling) and the old or initial schedule (before rescheduling).

3.0 RESCHEDULING METHODOLOGY

Three rescheduling approaches continuous, periodic and event-driven were defined by Ref. [1]. Continuous rescheduling takes a rescheduling action each time an event occurs. Periodic rescheduling defines rescheduling points between which any events that occur are ignored until the following rescheduling point. Finally, in the event-driven rescheduling, a rescheduling action is initiated upon an event with potential to cause significant disruption. Both continuous and periodic rescheduling can be viewed as special cases of event-driven rescheduling. In the ARSP, we take continuous and event-driven rescheduling approach where updating the existing schedule should take place when a rare event occurs. Rare events that have a potential to cause significant disruptions in the ARSP are interpreted by the arrival of new jobs, departure of an existing job, high deviations in the release times and changes in job priorities. Processing times and due date tightness are assumed fixed during the scheduling horizon.

There are generally three rescheduling repair methods: right shift rescheduling, partial rescheduling and regeneration. Right shift rescheduling postpones each remaining operation by the amount of time needed to make the schedule feasible. Partial rescheduling algorithm reschedules only the operations affected directly or indirectly by disruption. Regeneration reschedules the entire set of operations not processed before the rescheduling point, including those not affected by the disruption [1]. In the ARSP, a partial schedule repair procedure is developed to keep the unaffected part of the schedule and repair the later affected part by a dispatching rule.

The objective of the aerial refueling rescheduling problem is not only minimizing total weighted tardiness, but also minimizing schedule instability. Tardiness is defined as *max (0,* $C_j - d_j$*)* where C_j is completion time, d_j is the due date of job *j* and the total weighted tardiness is defined as $\sum w_j T_j$. Instability of the aerial refueling schedules is defined here as any changes in starting time on the assigned machine for each job. Then the measure of schedule instability can be defined as the proportion of rescheduled jobs that change machine assignment and starting time. Thus, in this paper, a heuristic algorithm that combines an appropriate dispatching rule for the weighted tardiness objective and partial repairing algorithm for schedule stability objective will be introduced.

3.1 General Partial Rescheduling Procedure

If we assume that one of the events occurs at time *t*, a general partial rescheduling procedure can be presented briefly as follows:

Step 1. Update weight and release time vectors (*W* and *R*) according to the event type and initialize processing time and due date input for old and new jobs
Step 2. Determine time *t'* to start repairing
Step 3. Determine *U* for time *t'*
Step 4. Repair the part of the schedule by assigning *U* using an appropriate dispatching rule
Step 5. Calculate objective function values according to the repaired schedule

U is the set of jobs that will be rescheduled point *t'*. *U* selection criterion to include jobs in the set of *U*, should keep stability of the current schedule meanwhile considering the weighted tardiness cost. The part of the existing schedule before *t'* will be kept as is to maintain schedule stability. In the ARSP, we use ATC rule priority index that will be explained in the next section, as a *U* selection criterion.

In Step 2, only weight (*W*) and release time (*R*) vectors are updated because any potential event type can be represented by an update in the weight and release time value of the job in the vector as given Table 1. *M* is a large number to represent presence or absence of a job in the scheduling environment. Due date vector is assumed to be defined as a function of processing time (*P*) and release time (R) vectors with fixed due date tightness, $D = R + \alpha.P$.

Table 1. Weight and release time changes for various event types

j : index of the job causing a disruption	Current Weight (*W*)	Updated Weight (*W'*)	Current Release Time (*R*)	Updated Release Time (*R'*)
Job Arrival	0	w_j	*M*	r_j
Job Departure	w_j	0	r_j	*M*
Changing Release Time	w_j	w_j	r_j	r_j'
Changing Priorities	w_j	w_j'	r_j	r_j

According to Table 1, if job arrival event occurs, *W* can be updated to *W'* by changing the zero weight value to assigned weight value (w_j) for the new job. Moreover, *R* can be updated to *R'* by changing *M* value (i.e. the corresponding job is not considered in the scheduling environment), to release time (r_j) for the new job. Changing release time and changing priority events can be illustrated by updating r_j to r_j' and w_j to w_j' respectively.

3.2 Apparent Tardiness Cost Rule

Dispatching (or Priority) Rules are the most common heuristics for scheduling problems due to their easy implementation and low computational power requirements. Apparent Tardiness Cost (ATC) heuristic is a good composite dispatching rule for the parallel machine total weighted tardiness problem. It combines the WSPT (Weighted Shortest Processing Time) rule and the Minimum Slack First rule (the job with the minimum slack is scheduled first) [4]. The priority index of ATC is defined as

$$\pi_j(\mathrm{t}) = \frac{w_j}{p_j}\exp\left[-\frac{\max\left[d_j - p_j - t, 0\right]}{k\overline{p}}\right] \qquad (1)$$

where

w_j : weight of the remaining job j

p_j : processing time of the remaining job j

d_j : due date of the remaining job j

t: Decision time point that the resource is considering which job to choose next.

$\bar{p}$: The average processing time of the remaining jobs,

k: 'look-ahead' or planning parameter and is set empirically (*k*= 2 is used here).

$S_j^+(t)$: the slack factor is equal to $max\ [d_j – p_j - t, 0]$.

This priority index is a function of the time t at which the machine become free. Under the ATC rule jobs are scheduled one at a time; that is, every time the machine becomes free the ranking index is computed for each remaining job. The job with the highest ranking index is then selected to be processed next.

4.0 PARTIAL RESCHEDULING ALGORITHM FOR JOB ARRIVAL DISRUPTION

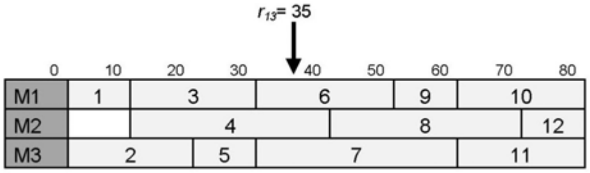
A partial rescheduling algorithm using the general procedure mentioned in Section 3 is developed for job arrival disruption and pseudo code of the algorithm is given as follows:

r_{new}: arrival time of the new job
r_j: release time of the job *j*
p_j: processing time of the job *j*
w_j weight of the job *j*
d_j: due date of the job *j*
$StartTime_j$: start time of the job *j*
$CompTime_j$: completion time of the job *j*
$Cmax_i$: makespan of the machine *i*
$\pi_j\,(t)$: priority of the job *j* at decision time *t*
S: set of jobs in the old schedule
U: set of waiting jobs to be rescheduled
C: set of jobs whose $\pi_j(t) > \pi_{new}(t)$
M: set of machines

```
1. Initialize $r_j$, $p_j$, $w_j$, $d_j$ $\forall j \in S$ and for the new job
2. Set $StartTime_j$ and $CompTime_j$ $\forall j \in S$ and
   $StartTime_{new}$ = large integer, $CompTime_{new}$ = large integer for the new job
3. Determine $Cmax_i$ = first completion time on machine $i$ after rnew, $\forall i \in M$
4. Set $t = \min_{i \in M} \{ Cmax_i \}$
5. Set $U=\{U \subseteq (S \cup \text{new job}) : StartTime_j \geq t \ \ \forall j \in U \}$
6. Calculate $\pi j(t)$ $\forall j \in U$ by using ATC
7. Determine set $C = \{ j \in C : \pi j(t) > \pi_{new}(t) \}$
8. Update $Cmax_i$ according to the latest $CompTime_j$ on machine $i$, $\forall i \in M$ and $j \in C$
9. if $StartTimej < Cmax_i$ $\forall j \in S$ and $\forall i \in M$ then remove $j$ from $U$
10. Update $t = \min_{i \in M}\{ Cmax_i \}$ ,
11. while $U \neq \emptyset$
12.       Find $j=\{j \in U : \pi j(t) = \max_{k \in U} \{\pi k\} \}$
13.       Find $i = \{i \in M : Cmax_i\ (t)= \min_{m \in M}\{ Cmax_m \} \}$
14.       Update $CompTime_j = Cmax_i + max(r_j,\ t) + p_j$ and remove $j$ from $U$
15.       Update $Cmax_i = CompTime_j$
16.       Update $t = \min_{i \in M}\{ Cmax_i \}$
17.       Calculate $\pi j(t)$ $\forall j \in U$ by using ATC
18. end while
19. Calculate and display the objective values, TWT and Instability.
```

The set *U* and start of the schedule repair are determined by steps 3-10. ATC rule priority index is used to determine the jobs in set *U* as a selection. The main rationale to determine set *U* is that jobs in set *C* (set of jobs whose $\pi_j(t) > \pi_{new}(t)$) cannot be scheduled later than new arriving job. Processes are assumed non-preemptive while determining decision time, *t*. Steps 11-18 is the repairing part of the algorithm that assigns the jobs to the machines by using ATC. In order to explain the partial rescheduling algorithm, a sample problem is given as follows :

A new job (the 13th job for the instance with number of jobs = 12, number of machines = 3) at time t = 35.



Step 1. Input data for an instance.

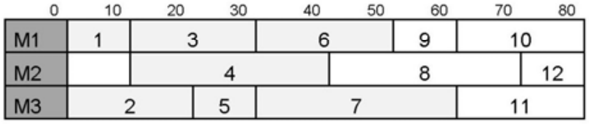
j	1	2	3	4	5	6	7	8	9	10	11	12	13
rj	0	0	10	10	20	30	30	40	50	60	60	70	35
pj	10	20	20	30	10	20	30	30	10	20	20	10	20
wj	8	7	5	4	8	5	3	9	8	9	6	1	5
dj	20	40	50	70	40	70	90	90	70	100	100	90	75

Step 2. Input start and completion times for set S and large integers for the new job.

job	1	2	3	4	5	6	7	8	9	10	11	12	13
StartTimej	0	0	10	10	20	30	30	40	50	60	60	70	M*
CompTimej	10	20	30	40	30	50	60	70	60	80	80	80	M*

M* = 99999

Step 3,4. Determine Cmaxi for i = {1,2,3} and t.



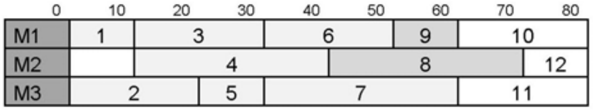
$Cmax_1$ = 50, $Cmax_2$ = 40, $Cmax_3$ = 60, t = min {50, 40, 60} = 40

Step 5. U = {8,9,10,11,12,13}

Step 6. $\pi_9(t)$= 0.609 > $\pi_8(t)$= 0.262 > $\pi_{13}(t)$=0.166 > $\pi_{11}(t)$= 0.152 > $\pi_{10}(t)$= 0.151 > $\pi_{12}(t)$= 0.051

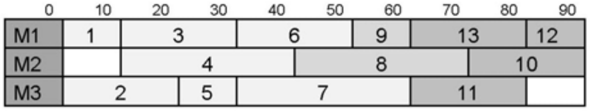
Step 7. C = {8,9} (jobs that have to be scheduled before the new job).

Step 8,9,10. $CompTime_8$ = 60 and $CompTime_9$ = 70



$Cmax_1$ = 60, $Cmax_2$ = 70, $Cmax_3$ = 60, U = {10,11,12,13}, t = min {60, 70, 60} = 60

Step 11-18.



Step 19. TWT = 70, $Instability$ = 3 (Jobs 10, 11, 12)

5.0 COMPUTATIONAL STUDY

To measure the effectiveness of the partial rescheduling algorithm, it is compared with the complete (regeneration) rescheduling algorithm in which all jobs after the arrival the new job are rescheduled using ATC rule similar to the way described in Steps 11-18 of the partial rescheduling algorithm. For this aim, the TWT and Instability (number of jobs whose start time or assigned machine has changed) objective values were obtained for different combinations of release times (10, 30, 50), processing times (15, 30, 45) and weights (1, 5, 9) of the arriving job. Experimental results are given in Table 2. Due dates are generated by equation $d_j = r_j + \alpha.p_j$ with $\alpha = 2$.

If we assume that instance samples were drawn independently and randomly from the instance population for two algorithms, the single factor ANOVA for comparing the algorithms in TWT shows that the partial rescheduling is superior to the complete rescheduling in terms of TWT at 80% significance level. ANOVA was also performed on the Instability and the partial rescheduling is significantly superior to the complete algorithm.

Table 2. TWT and Instability Values for Partial and Complete Rescheduling

Instance	Release Time	Processing Time	Weight	Total Weighted Tardiness		Instability*	
				Partial Rescheduling	Complete Rescheduling	Partial Rescheduling	Complete Rescheduling
1	10	15	1	261.6	301.7	3	10
2	10	15	5	355.2	371.6	6	10
3	10	15	9	237.7	237.7	10	10
4	10	30	1	208.6	309.5	2	10
5	10	30	5	426.5	461.1	4	10
6	10	30	9	418.9	411.0	6	10
7	10	45	1	193.6	294.5	0	10
8	10	45	5	359.2	470.1	3	10
9	10	45	9	404.5	502.5	4	10
10	30	15	1	190.0	225.5	2	8
11	30	15	5	255.2	232.5	6	8
12	30	15	9	233.5	233.5	8	8
13	30	30	1	188.6	215.0	0	8
14	30	30	5	326.5	272.6	4	8
15	30	30	9	276.6	310.9	6	8
16	30	45	1	173.6	200.0	0	8
17	30	45	5	248.8	301.6	2	8
18	30	45	9	349.7	362.6	4	8
19	50	15	1	170.0	199.0	2	6
20	50	15	5	201.2	201.2	6	6
21	50	15	9	201.2	201.2	6	6
22	50	30	1	168.6	189.7	0	6
23	50	30	5	222.2	292.6	3	6
24	50	30	9	321.1	279.2	4	6
25	50	45	1	153.6	174.7	0	6
26	50	45	5	154.8	191.1	2	6
27	50	45	9	237.2	369.2	3	6

* Instability : Number of jobs whose start time or assigned machine has changed.

6.0 CONCLUSIONS

This paper presents a partial rescheduling algorithm for job arrival disruptions in the parallel machine scheduling. A heuristic algorithm that combines an appropriate dispatching rule (the ATC) for the total weighted tardiness objective with partial repair algorithm for the schedule stability objective, was introduced. This algorithm can be modified to use for other types of job related disruptions. To measure the effectiveness of the partial rescheduling algorithm, it was compared with the complete (regeneration) rescheduling algorithm. According to ANOVA, the partial rescheduling has superior results compared to the complete rescheduling. In future research, more constraints such as machine compatibility, sequence dependent setup times and deadlines may be included in the model. In addition, insertion algorithms that emphasize more the instability objective may be developed. Other composite dispatching rules and metaheuristics may be implemented to obtain better results and linear models may be programmed to find optimal solutions that can be used for comparisons.

7.0 REFERENCES

[1] Church, L.K., and Uzsoy, R., Analysis of periodic and event driven rescheduling policies in dynamic shops, International Journal of Computer Integrated Manufacturing, 5, 153–163, 1992.

[2] Curry, J. and Peters, B., Rescheduling parallel machines with stepwise increasing tardiness and machine assignment stability objectives, International Journal of Production Research,43:15, 3231 – 3246, 2005.

[3] Duenas, A., and Petrovic, D., An approach to predictive-reactive scheduling of parallel machines subject to disruptions, Annual Operations Research, 159, 65–82, 2008.

[4] Pinedo, M., Scheduling theory, algorithms, and systems. 3rd Ed., Springer, New York, 2008.

Aerial Refueling Process Rescheduling Under Job Related Disruptions

Sezgin KAPLAN, Ph.D. Student
Old Dominion University

October 13, 2010

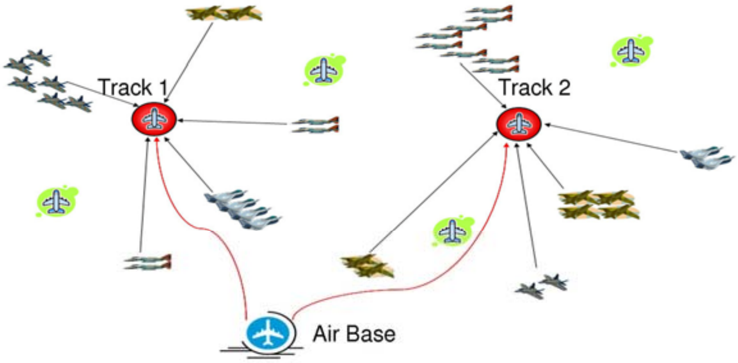
Outline

- Aerial Refueling Scheduling Problem (ARSP)
- Rescheduling Problem
- AR Rescheduling Problem
- Partial Rescheduling Algorithm
- Apparent Tardiness Cost (ATC) Rule
- An Example
- Computational Study
- Conclusion
- References

ARSP

- Aerial Refueling: The process of transferring fuel from one aircraft (a tanker) to another (a fighter as a receiver) during flight.

ARSP



- ARSP: Determining the refueling completion times for each fighter aircraft (job) on one of multiple tankers (machines).

ARSP

- **Model :** An identical parallel machine scheduling problem with release times and due dates.
- A system with *m* identical machines in parallel and *n* jobs where job *j* becomes available at ready time r_j and should be complete and leave by the due date d_j.
- **Objective :** To minimize total weighted tardiness (TWT) as a performance measure.

Rescheduling Problem

- Disruptions caused by dynamic and unexpected events require rescheduling to update the existing aerial refueling schedule.
- **Types of events :** New (rush) job arrivals, order cancellations, machine failure, changes in order priority/ready times/processing times/due dates, rework due to quality problems, material/operator/tool unavailability etc [1,2,3,4].

Rescheduling Problem

- Continuous, periodic and event-driven rescheduling approaches [1].
- Right shift rescheduling, partial rescheduling and regeneration methods.
- **Multi objectives :** objective of the original problem + minimization of the difference between old and new schedules.

AR Rescheduling Problem

- Job related disruptions :
  - the arrival of new jobs,
  - the departure of an existing job,
  - high deviations in the release times,
  - changes in job priorities.
- Continuous and event-driven rescheduling approach.
- Partial schedule repair method.

MODSIM WORLD
Conference & Expo

AR Rescheduling Problem

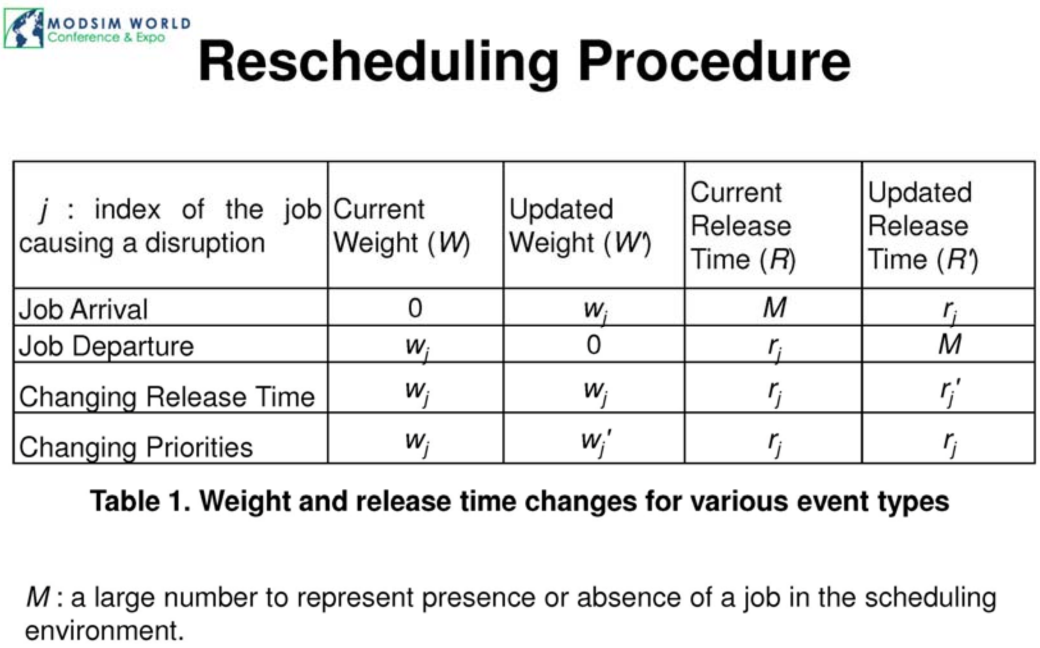
- **Objective :** minimize TWT + minimize instability.
- **Tardiness :** *max(0,C_j -d_j)* where C_j is completion time, d_j is the due date of job *j.*
- Total weighted tardiness (TWT) : $\sum w_j T_{j.}$
- **Instability :** any changes in starting time on the assigned machine for each job.
- **Measure of instability :** the proportion of rescheduled jobs that change machine assignment and starting time.

MODSIM WORLD
Conference & Expo

Partial Rescheduling Algorithm

- General partial rescheduling procedure can be used for different types of job related disruptions.
  1. Initialize data (ready time, processing time, weight, due date) according to the event type occurred at time *t.*
  2. Determine the part of the schedule to repair.
  3. Repair the part by using an appropriate dispatching rule
- A heuristic algorithm for arrival of new jobs :

  Apparent Tardiness Cost (ATC) dispatching rule + Partial repair algorithm.

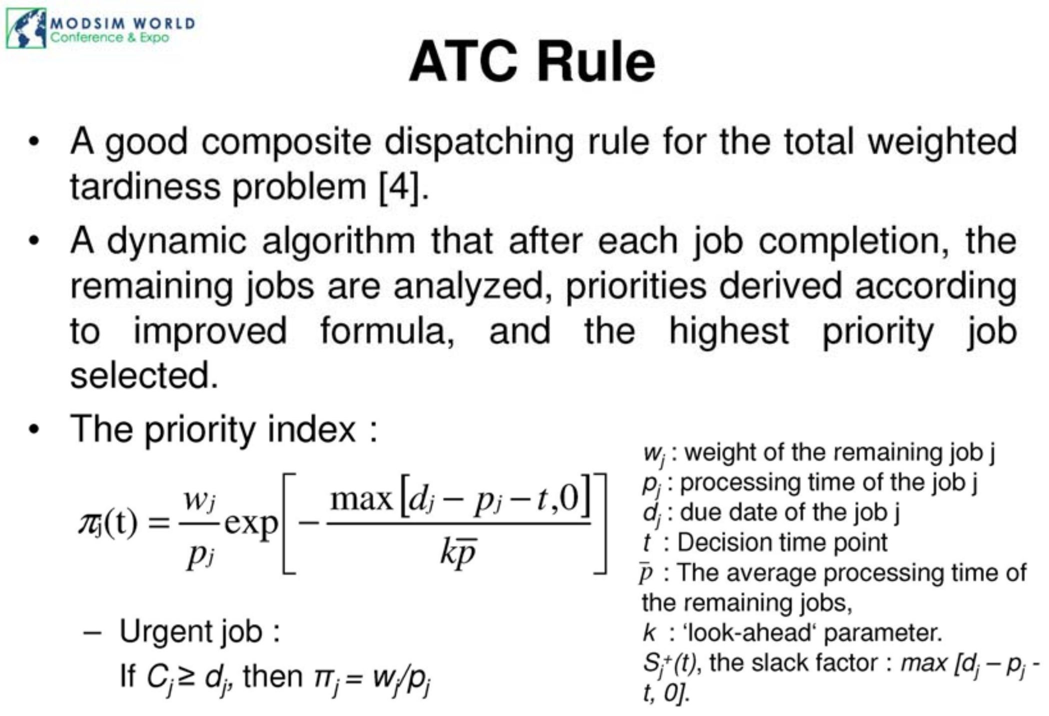


Rescheduling Procedure

j : index of the job causing a disruption	Current Weight (W)	Updated Weight (W')	Current Release Time (R)	Updated Release Time (R')
Job Arrival	0	w_j	M	r_j
Job Departure	w_j	0	r_j	M
Changing Release Time	w_j	w_j	r_j	r_j'
Changing Priorities	w_j	w_j'	r_j	r_j

Table 1. Weight and release time changes for various event types

M : a large number to represent presence or absence of a job in the scheduling environment.



ATC Rule

- A good composite dispatching rule for the total weighted tardiness problem [4].
- A dynamic algorithm that after each job completion, the remaining jobs are analyzed, priorities derived according to improved formula, and the highest priority job selected.
- The priority index :

$$\pi_j(t) = \frac{w_j}{p_j}\exp\left[-\frac{\max\left[d_j - p_j - t, 0\right]}{k\overline{p}}\right]$$

  - Urgent job :
    If $C_j \geq d_j$, then $\pi_j = w_j/p_j$

w_j : weight of the remaining job j
p_j : processing time of the job j
d_j : due date of the job j
t : Decision time point
$\overline{p}$: The average processing time of the remaining jobs,
k : 'look-ahead' parameter.
$S_j^+(t)$, the slack factor : $max\ [d_j - p_j - t, 0]$.

Partial Rescheduling Algorithm

r_{new}: arrival time of the new job
r_j: release time of the job j
p_j: processing time of the job j
w_j weight of the job j
d_j: due date of the job j
$StartTime_j$: start time of the job j
$CompTime_j$: completion time of the job j
$Cmax_i$: makespan of the machine i
$\pi_j(t)$: priority of the job j at decision time t
S: set of jobs in the old schedule
U: set of waiting jobs to be rescheduled
C: set of jobs whose $\pi_j(t) > \pi_{new}(t)$
M: set of machines

```
1.  Initialize $r_j, p_j, w_j, d_j$ $\forall j \in S$ and for the new job
2.  Set $StartTime_i$ and $CompTime_i$ $\forall j \in S$ and
    $StartTime_{new}$ = large integer, $CompTime_{new}$ = large integer for the new job
3.  Determine $Cmax_i$ = first completion time on machine $i$ after mew, $\forall i \in M$
4.  Set $t = \min_{i \in M} \{ Cmax_i \}$
5.  Set $U = \{U \subseteq (S \cup$ new job) : $StartTime_i \geq t \ \ \forall j \in U \}$
6.  Calculate $\pi j(t)$ $\forall j \in U$ by using ATC
7.  Determine set $C = \{ j \in C : \pi j(t) > \pi_{new}(t) \}$
8.  Update $Cmax_i$ according to the latest $CompTime_i$ on machine $i$, $\forall i \in M$ and $j \in C$
9.  if $StartTimej < Cmax_i$ $\forall j \in S$ and $\forall i \in M$ then remove $j$ from $U$
10. Update $t = \min_{i \in M} \{ Cmax_i \}$ .
11. while $U \neq \emptyset$
12.     Find
        j=$\{j \in U : \pi j(t) = \max_{k \in U} \{\pi k\} \}$
13.     Find $i = \{i \in M : Cmax_i(t) = \min_{m \in M} \{ Cmax_m \} \}$
14.     Update $CompTime_i = Cmax_i + max(r_i, t) + p_i$ and remove $j$ from $U$
15.     Update $Cmax_i = CompTime_i$
16.     Update $t = \min_{i \in M} \{ Cmax_i \}$
17.     Calculate $\pi j(t)$ $\forall j \in U$ by using ATC
18. end while
19. Calculate and display the objective values, TWT and Instability.
```

Input : Steps 1 and 2

Determine the *U* set to reschedule: Steps 3-10

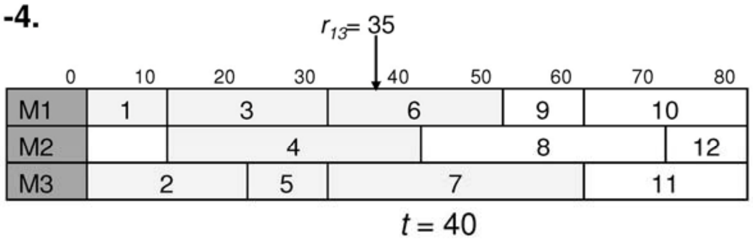
The main rationale : jobs in set *C* (set of jobs whose $\pi_j(t) > \pi_{new}(t)$) cannot be scheduled later than new arriving job.

Reschedule by ATC : Steps 11-18

Display

An Example

Step 1-4.

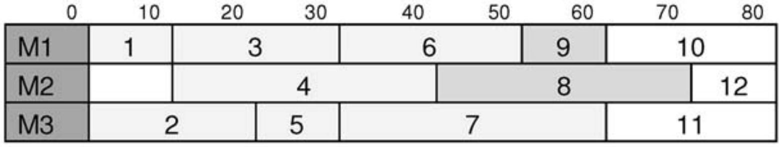


Step 5. U = {8,9,10,11,12,13}

Step 6. $\boldsymbol{\pi_9(40)}$**= 0.609 >** $\boldsymbol{\pi_8(40)}$**= 0.262** > $\pi_{13}(40)$=0.166 > $\pi_{11}(40)$= 0.152 > $\pi_{10}(40)$= 0.151 > $\pi_{12}(40)$= 0.051

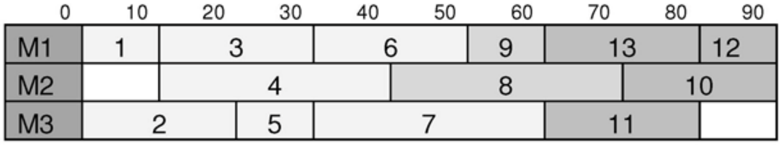
Step 7. C = {8,9} (jobs that have to be scheduled before the new job).

An Example



Step 8-10. *CompTime*$_8$ = 60 and *CompTime*$_9$ = 70
Cmax$_1$ = 60, *Cmax*$_2$ = 70, *Cmax*$_3$ = 60,
U = {10,11,12,13}, *t* = *min {60, 70, 60}* = 60

Step 11-18. $\pi_{13}(60) > \pi_{11}(60) > \pi_{10}(60) > \pi_{12}(60)$



Step 19. *TWT* = 70, *Instability* = 2 (Jobs 10 and 12)

Computational Study

Table 2. TWT and Instability Values for Partial and Complete Rescheduling

Instance	Release Time	Processing Time	Weight	Total Weighted Tardiness		Instability*	
				Partial Rescheduling	Complete Rescheduling	Partial Rescheduling	Complete Rescheduling
1	10	15	1	261.6	301.7	3	10
2	10	15	5	355.2	371.6	6	10
3	10	15	9	237.7	237.7	10	10
4	10	30	1	208.6	309.5	2	10
5	10	30	5	426.5	461.1	4	10
6	10	30	9	418.9	411.0	6	10
7	10	45	1	193.6	294.5	0	10
8	10	45	5	359.2	470.1	3	10
9	10	45	9	404.5	502.5	4	10
10	30	15	1	190.0	225.5	2	8
11	30	15	5	255.2	232.5	6	8
12	30	15	9	233.5	233.5	8	8
13	30	30	1	188.6	215.0	0	8
14	30	30	5	326.5	272.6	4	8
15	30	30	9	276.6	310.9	6	8
16	30	45	1	173.6	200.0	0	8
17	30	45	5	248.8	301.6	2	8
18	30	45	9	349.7	362.6	4	8
19	50	15	1	170.0	199.0	2	6
20	50	15	5	201.2	201.2	6	6
21	50	15	9	201.2	201.2	6	6
22	50	30	1	168.6	189.7	0	6
23	50	30	5	222.2	292.6	3	6
24	50	30	9	321.1	279.2	4	6
25	50	45	1	153.6	174.7	0	6
26	50	45	5	154.8	191.1	2	6
27	50	45	9	237.2	369.2	3	6

* Instability : Number of jobs whose start time or assigned machine has changed.

Complete (regeneration) rescheduling algorithm : All jobs after the arrival of the new job are rescheduled using ATC rule.

Conclusion

- A heuristic algorithm that combines an appropriate dispatching rule (the ATC) for the total weighted tardiness objective with partial repair algorithm for the schedule stability objective, was introduced.
- To measure the effectiveness of the partial rescheduling algorithm, it was compared with the complete (regeneration) rescheduling algorithm. The partial rescheduling has superior results compared to the complete rescheduling.
- In future research, more constraints such as machine compatibility, sequence dependent setup times and deadlines may be included in the model. In addition, insertion algorithms that emphasize more the instability objective may be developed. Other composite dispatching rules and metaheuristics may be implemented to obtain better results.

References

[1] Church, L.K., and Uzsoy, R., Analysis of periodic and event driven rescheduling policies in dynamic shops, International Journal of Computer Integrated Manufacturing, 5, 153–163, 1992.

[2] Curry, J. and Peters, B., Rescheduling parallel machines with stepwise increasing tardiness and machine assignment stability objectives, International Journal of Production Research,43:15, 3231 – 3246, 2005.

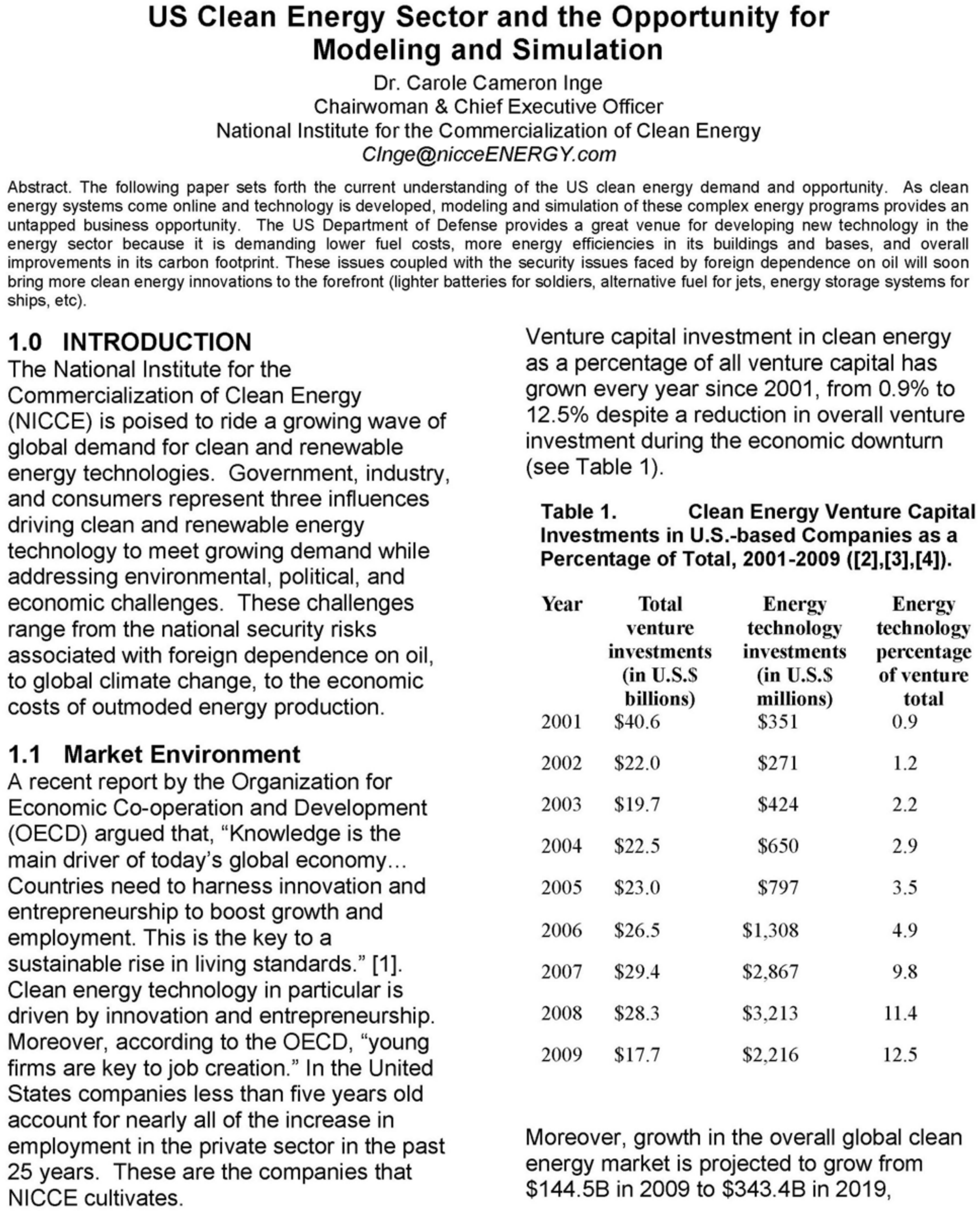
[3] Duenas, A., and Petrovic, D., An approach to predictive-reactive scheduling of parallel machines subject to disruptions, Annual Operations Research, 159, 65–82, 2008.

[4] Pinedo, M., Scheduling theory, algorithms, and systems. 3rd Ed., Springer, New York, 2008.

MODSIM WORLD
Conference & Expo

Questions/Comments

2.9 US Clean Energy Sector and the Opportunity for Modeling and Simulation



US Clean Energy Sector and the Opportunity for Modeling and Simulation

Dr. Carole Cameron Inge
Chairwoman & Chief Executive Officer
National Institute for the Commercialization of Clean Energy
CInge@nicceENERGY.com

Abstract. The following paper sets forth the current understanding of the US clean energy demand and opportunity. As clean energy systems come online and technology is developed, modeling and simulation of these complex energy programs provides an untapped business opportunity. The US Department of Defense provides a great venue for developing new technology in the energy sector because it is demanding lower fuel costs, more energy efficiencies in its buildings and bases, and overall improvements in its carbon footprint. These issues coupled with the security issues faced by foreign dependence on oil will soon bring more clean energy innovations to the forefront (lighter batteries for soldiers, alternative fuel for jets, energy storage systems for ships, etc).

1.0 INTRODUCTION

The National Institute for the Commercialization of Clean Energy (NICCE) is poised to ride a growing wave of global demand for clean and renewable energy technologies. Government, industry, and consumers represent three influences driving clean and renewable energy technology to meet growing demand while addressing environmental, political, and economic challenges. These challenges range from the national security risks associated with foreign dependence on oil, to global climate change, to the economic costs of outmoded energy production.

1.1 Market Environment

A recent report by the Organization for Economic Co-operation and Development (OECD) argued that, "Knowledge is the main driver of today's global economy… Countries need to harness innovation and entrepreneurship to boost growth and employment. This is the key to a sustainable rise in living standards." [1]. Clean energy technology in particular is driven by innovation and entrepreneurship. Moreover, according to the OECD, "young firms are key to job creation." In the United States companies less than five years old account for nearly all of the increase in employment in the private sector in the past 25 years. These are the companies that NICCE cultivates.

Venture capital investment in clean energy as a percentage of all venture capital has grown every year since 2001, from 0.9% to 12.5% despite a reduction in overall venture investment during the economic downturn (see Table 1).

Table 1. Clean Energy Venture Capital Investments in U.S.-based Companies as a Percentage of Total, 2001-2009 ([2],[3],[4]).

Year	Total venture investments (in U.S.$ billions)	Energy technology investments (in U.S.$ millions)	Energy technology percentage of venture total
2001	$40.6	$351	0.9
2002	$22.0	$271	1.2
2003	$19.7	$424	2.2
2004	$22.5	$650	2.9
2005	$23.0	$797	3.5
2006	$26.5	$1,308	4.9
2007	$29.4	$2,867	9.8
2008	$28.3	$3,213	11.4
2009	$17.7	$2,216	12.5

Moreover, growth in the overall global clean energy market is projected to grow from $144.5B in 2009 to $343.4B in 2019, distributed fairly evenly among biofuels, wind power, and solar power [2].

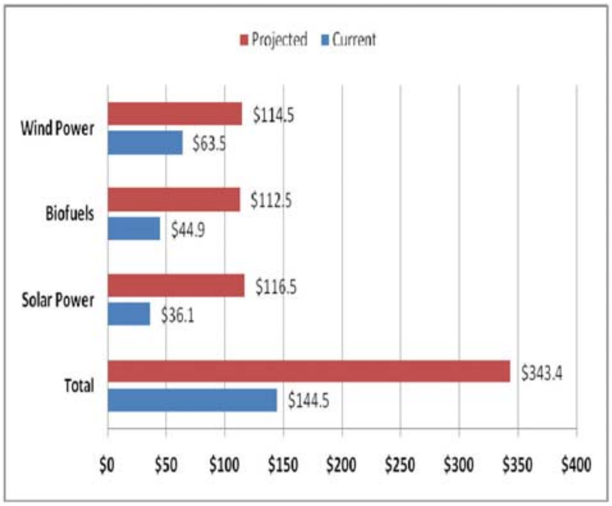


Figure 1. Global Clean-Energy Projected Growth 2009-2019 [2].

The clean energy market is forecast to expand by more than 200% in the next 10 years alone. This rapid growth will be driven by three categories of technology identified by the National Renewable Energy Laboratory (NREL), the leading national laboratory for alternative and renewable energy under the US Department of Energy:

- Accelerated evolutionary technologies driven by industry, approximately 3 years from market
- Disruptive technologies driven by major technological advances, approximately 3-10 years from market
- Revolutionary technologies driven by basic research, 10 years or more from market.

The projected market growth shown below will be almost entirely the result of accelerated evolutionary and disruptive technologies. NICCE is aimed precisely at companies positioned in these two stages.

1.2 Government Incentives

The United States Department of Energy (DoE), Department of Defense (DoD), and Department of Agriculture (USDA)—the three most notable governmental departments in terms of funding and innovation—all have funding and mandates to develop clean energy technologies. The NREL summarized the 2010 market environment from the government perspective:

1. Invest $150B in alternative energy over 10 years.

2. Create green jobs with clean, efficient American energy.

3. Double production of alternative energy in three years– enough to power 6 million homes.

4. Upgrade the efficiency of more than 75% of federal buildings and two million private homes.

5. Put one million PHEVs (Plug-in Hybrid Electric Vehicles) on U.S. roads by 2015.

6. Reduce CO2 emissions by 80% below 1990 levels by 2050.

7. Transform the economy with science and technology.

For NICCE, this commitment is an opportunity, as federal, state, and local governments are looking for viable clean and renewable technologies ready for commercialization in the short term—3 to 10 years at most.

2.0 MAJOR DRIVERS

The following sections detail major drivers behind governmental and industry demand for accelerated evolutionary and disruptive clean and renewable energy technology in the next 10 years, including government regulations and mandates, industry investment in clean energy research and development, and consumer interest and advocacy.

2.1 US Department of Energy

The Department of Energy has detailed Energy Management Requirements that capture various legislative initiatives including Executive Order (E.O.) 13514, the Energy Independence and Security Act of 2007 (EISA 2007), E.O. 13423, the Energy Policy Act of 2005 (EPAct 2005), the National Energy Conservation Policy Act (NECPA), and other policies.

Federal agencies must meet the energy management requirements outlined by federal statutory laws and regulations. The following clean energy technologies drawn from the Department of Energy summary of Energy Management requirements, which apply across multiple departments, represent areas of opportunity for NICCE:

2.2 Renewable Energy-Use Increase

The EPAct 2005 defines "renewable energy" as electric energy generated from solar, wind, biomass, landfill gas, ocean (including tidal, wave, current, and thermal), geothermal, municipal solid waste, or new hydroelectric generation capacity achieved from increased efficiency or additions of new capacity at an existing hydroelectric project.

It requires that the total electricity consumed by the federal government coming from renewable energy is:

- Not less than 3% in fiscal years 2007-2009
- Not less than 5% in fiscal years 2010-2012
- Not less than 7.5% in fiscal year 2013 and thereafter

Similarly, E.O. 13423 mandates that at least half of renewable energy used by the federal government must come from new renewable sources (in service after January 1, 1999). Again, this suggests an opportunity for NICCE since federal agencies will be searching for new clean and renewable sources.

2.3 Petroleum Use Reduction/Alternative Fuel-Use Increase

The EISA 2007 requires federal agencies to achieve a 20% reduction in petroleum consumption by 2015 compared to a fiscal year 2005 baseline. Moreover, it requires federal agencies to increase alternative fuel use 10% each year compared to a fiscal year 2005 baseline. E.O. 13423 requires federal agencies with 20 vehicles or more located in the U.S. to decrease petroleum consumption by 2% per year through fiscal year 2015 compared to a fiscal year 2005 baseline and requires federal agencies to increase alternative fuel use by 10% each year compared to the previous year. Both of these measures provide an opportunity for biofuels production, a key element of the NICCE market.

2.4 US Department of Defense

A more specific example comes from the Department of Defense, the single largest consumer of energy in the United States. For the DoD, clean energy technology falls at the intersection of national security, economic security, and environmental security.

Existing DoD consumption is heavily skewed towards jet fuel and electricity, providing a distinctive opportunity for biofuels and renewable electricity generation, including "smart grid" technology, both areas of opportunity for NICCE as identified by our commissioned Research Triangle International (RTI) report on clean energy incubators (see Figure 2).

According to the DoD 2010 Quadrennial Defense Review Report, "The Department is increasing its use of renewable energy supplies and reducing energy demand to improve operational effectiveness, reduce greenhouse gas emissions in support of U.S. climate change initiatives, and protect the Department from energy price

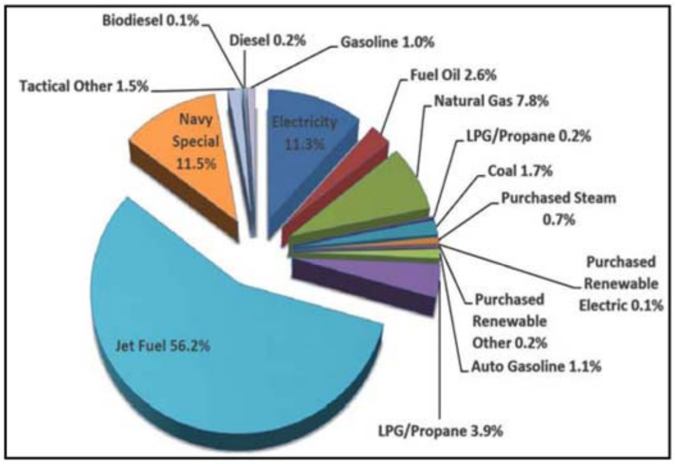


Figure 2. Existing DOE Energy Consumption [5].

fluctuations." This is not empty rhetoric. Significant directives mandating specific energy performance targets as shown in Table 2 back the report.

Table 2. DoD Energy Performance Targets [5].

Area	Goal	Legislation
Installations energy use	Reduce by 30% by 2015 from 2003 baseline.	EO 13423 / EISA 2007
Non-tactical vehicle (NTV) fuel consumption	Reduce 2% per year through 2015, 20% by 2015.	EO 13423
Electricity from renewable sources	A "Sense of Congress" goal: Reduce 25% by 2025.	EISA 2007 / NDAA 2007
Fossil fuel use in new/renovated buildings	Reduce 55% by 2010; 100% by 2030	EISA 2007
Hot water in new/renovated buildings from solar power	30% by 2015 if life-cycle is cost-effective	EISA 2007
Non-petroleum fueled vehicles use (ethanol, natural gas)	Increase by 10% annually.	EO 13423
Energy metering for improved energy management	Meter electricity by Oct 2012	EPAct200
	Meter natural gas and steam by October of 2016.	EISA 2007

2.5 US Department of Agriculture

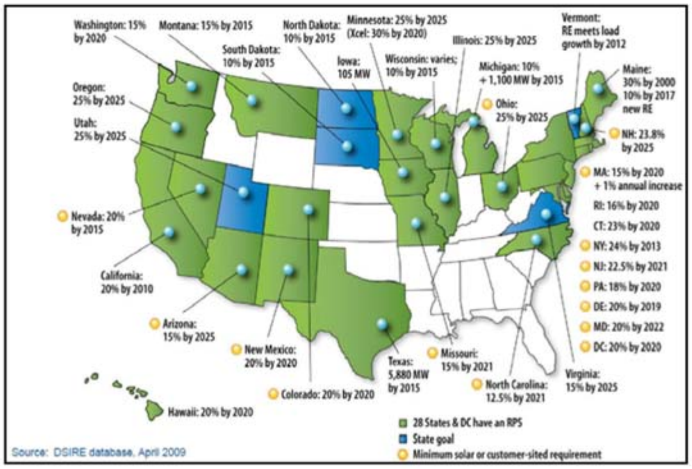
The USDA's annual budget of $95B, includes major support programs for rural renewable energy projects and direct support for biofuels such as ethanol and biodiesel. The 2008 USDA Energy Council, report "Advancing Renewable Energy" highlights the Renewable Fuels Mandate, which requires a 500% increase in the use of renewable fuels to 36 billion gallons annually by year 2022. Similarly, the Federal Government Operations Mandate calls for a 30% reduction in energy consumption by federal government facilities by 2015.

Through its Rural Development grant and loan programs, the USDA implements commercialization strategies and supports agriculture producers and forest landowners, rural small businesses, electric cooperatives, and other rural investors in deploying renewable technologies such as ethanol, biodiesel, methane gas recovery, and wind, solar, and geothermal power. Moreover, USDA has demonstrated a commitment to the market adoption of renewable energy technologies as part of the mainstream energy grid. From 2001 through 2007, more than $674 million in USDA funds were distributed to 1,763 renewable energy research, economic development, and energy efficiency initiatives. These investments translated to an 80.3 million metric ton reduction of CO2 emissions and a savings/production of approximately 2.4 billion kilowatt hours of energy. In 2007, USDA committed nearly $75 million toward renewable energy programs, including research and development of cellulosic ethanol—a form of ethanol fuel created from switch grass, wood chips, and other woody biomass.

In recent years, NICCE and the USDA have partnered for the creation of green growth in rural areas creating jobs in local communities such as Southern Virginia.

2.6 State & Local Government Programs

While federal demand for clean and renewable energy is the main market driver in the clean energy economy, State and Local governments often control federal funds as well as local facilities, tax incentives, and other supports. Understanding those local structures is vital in gaining access to federal and state stimulus funding and tapping into the desire to attract "green collar" jobs to communities. These state and local communities represent distributed drivers of the clean energy economy. NREL has identified the "renewable portfolio standards" (RPS), regulations that require increased production of energy from renewable energy sources such as wind, solar, biomass, and geothermal for states around the country, as well as highlighting state renewable energy goals and solar-specific goals. The map below shows RPS goals.



Source: DSIRE database, April 2009

Figure 3. NREL Renewable Portfolio Standards [6]

In summary, by combining federal, state, and local demand, NICCE leverages all levels of government support, providing a wider network and reaching communities around the country while maintaining its national presence.

2.7 Industry

For industry, clean and renewable energy technology is becoming both economically and politically necessary as energy prices rise, governments implement more rigorous environmental standards, and consumers exert greater demand for clean, sustainable, and responsible business. While total investment in new clean energy generation capacity has outpaced conventional fossil fuels for the past two years in a row, the recent economic downturn has resulted in major reductions in venture capital investments, including in clean energy technology. Yet at the same time, investment in late-stage companies increased by 19%. Investment companies are making a clear move towards companies on the cusp of "crossing the chasm," exactly the point at which NICCE offers a distinctive commercialization process.

2.8 Consumer

Consumer demand for affordable clean energy technology is unequivocal. Energy companies are scrambling to clean up their brands. The recent BP oil spill only serves to reinforce public perceptions. However, the current economic climate-- from unemployment to foreclosures-- has consumers wary of rising energy costs. Clean energy will only be appealing to consumers if it is affordable, reliable, and convenient. The NICCE commercialization process captures this demand, focusing on economic and technological realities in order to evaluate potential clean energy technologies and identify markets.

3.0 DISCUSSION – NICCE BUSINESS MODEL

NICCE is building the Modeling and Simulation Center for Excellence near Washington, DC, in the Dulles Corridor and plans to have satellite facilities throughout the world that bring together inventors, investors, and local communities to enable the commercialization of clean energy technologies and the creation of "green collar" jobs around the United States.

NICCE identifies and cultivates entrepreneurial clean energy companies, taking them through a proven commercialization process based on strategic alliances with professional engineering and business service providers, modeling and simulation centers such as VMASC, advanced research laboratories, and potential investors. Companies progress through the NICCE commercialization framework with support from local NICCE-approved community-based clean energy business incubators. As these businesses mature, NICCE connects them with governmental, institutional, and private investors as well as manufacturing parties to bring secure, affordable, and clean energy technologies to market.

3.1 Next Generation Clean Tech Incubators

Historically, community-based business incubators target multiple industries and offer an array of generalized services to the companies within them to produce successful firms that will graduate from the program financially viable and freestanding. These incubator graduates have the potential to create jobs, revitalize neighborhoods, commercialize new technologies, and strengthen local and national economies." [7].

To reach this goal, incubators typically provide a set of services to help these companies reach self-sustainability and realize their market potential. With the help of NICCE, increasing numbers of organizations are going beyond this traditional set of incubation services to provide highly specialized technology business services, linking incubated companies with key R&D assets (national energy labs, universities, etc), and even providing capital investment. These organizations frequently have an economic development mission, and focus on a single industry or relatively narrow set of technology areas.

NICCE advances its model of enhanced incubator services further by remaining dynamic and flexible, tailoring its services and its own development trajectory to the rapidly evolving nature of the clean energy industry. With a better understanding of clean energy development, NICCE recognizes the need for sophisticated services to include modeling and simulation, third party energy validation, remote sensing services, a World Data Center on Energy, strategic finance alliances, product optimization, energy policy and government affairs, sales, marketing and business development augmentation. According to Research Triangle International, a group commissioned by NICCE, this new next generation incubator model is revolutionary and on target with the growing needs of the clean energy industry. Therefore, NICCE is developing a mega incubator called the National Capital Clean Energy Incubator to provide other incubators and their respective companies these services.

NICCE's national and international brand is sleek, accessible, and affordable with high ROI for its investors. NICCE's current customers believe that this model creates a value-added supply chain that is missing in the clean industry commercialization process today. These customers include semiconductor companies in smart grid, biofuels customers and grid watt level energy storage customers, to name a few.

3.2 Clean Energy Commercialization

While technology commercialization in general is well understood, clean energy commercialization poses a number of unique challenges. Extensive experience has led the NICCE management team to develop the following process for commercialization. Companies entering NICCE go through an extensive evaluation process to determine their position in the market and the services they need. In working with local NICCE approved incubators and through the NICCE strategic partnerships businesses get professional services support, modeling and simulation, and other programs to help the companies cross the "chasm" and ultimately get to market.

NICCE management believes that it is uniquely positioned to provide commercialization services along the business development continuum, with an emphasis on identifying mid-stage companies that still require development but present a considerable investment opportunity. This mid-stage approach is captured well by a concept proposed by Geoffrey Moore in his book, "Crossing the Chasm: Marketing and Selling High-Tech Products to Mainstream Customers." [8].

A recurring problem with new technology companies occurs after early adopters have proven the concept but before there is large-scale adoption. This marketing, development and financing gap always looms large in any new technology company's future. NICCE gets companies to the chasm and helps them across through partnerships, large-scale government contracts, and its investor relationships. NICCE evaluates whether companies are likely to succeed in the jump across "the chasm" and then helps them make that jump through the NICCE clean energy commercialization process (see Figure 4).

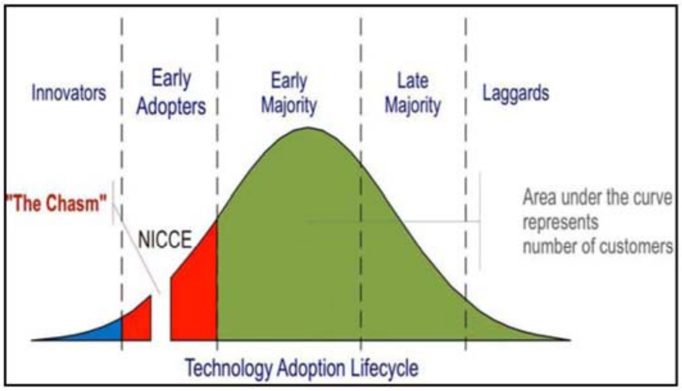


Figure 4. **NICCE Commercialization Process.**

NICCE's commercialization process begins with an evaluation—where does a given company sit on the commercialization spectrum? This allows NICCE to place the company at a given point along the process and identify the services and steps necessary to take that company to commercial success. The flow chart that follows illustrates that process.

The initial assessment is key in this process, as it establishes whether a company is ready to enter NICCE and whether it has any potential for commercialization. Three key questions NICCE asks during this phase are:

1. Is the technology best in class?

2. Can the company bring the technology to market within 36 months?

3. Is there a large-scale customer, like the DoD, a utility or NASA and can the company meet volume and profit targets?

By emphasizing this "entry test," NICCE can assure investors of validated opportunities while earning a reputation among start-ups for honest evaluation. This saves both inventors and investors time and money. Moreover, NICCE's sometimes takes an equity stake in member companies and it aligns the interests of all three parties—the success of a NICCE member translates into success for NICCE, its strategic partners, and its investors.

The commercialization process itself is the result of years of experience on behalf of the management team in starting, cultivating, and commercializing technologies and businesses. Each step is accompanied by a comprehensive package of services that experience has shown help a given company "cross the chasm" to commercialization and success.

4.0 CONCLUSION(S)

In conclusion, NICCE is leading the way toward a greener US economy. The models developed herein represent best practices for clean energy commercialization and these practices ultimately will lead to job growth in local communities. Modeling and simulation, along with the World Data Center on Energy represent a value added service and much needed asset for all clean energy development. The private sector, along with national laboratories and universities must work hand in hand to grow this next generation industry. NICCE serves as a catalyst for these groups coming together to grow and develop a new wave of innovation in the US.

5.0 REFERENCES

[1] OECD (2010), "The OECD Innovation Strategy: Key Findings", document prepared for the 2010 meeting of the OECD Council at Ministerial level, Paris, 27-28 May 2010.

[2] Clean Edge, "Clean Energy Trends 2010". March 2010.

[3] Bloomberg New Energy Finance, September 2010, http://bnef.com/

[4] Nth Power, Innovating for Better Buildings, http://www.nthpower.com/

[5] Global Green, USA. Defense Sustainability: Energy Efficiency and the Battlefield. Schuyler Null. February 2010.

[6] Geiss, Kevin, "Army Energy Security Presentation," U.S. Army (October 19, 2009). National Renewable Energy Lab maps and National Geospatial Intelligence Agency data services, ESRI Data & Maps CD. Created in ArcGIS 9.3 using Ardnfo (Concurrent Technologies Corporation).

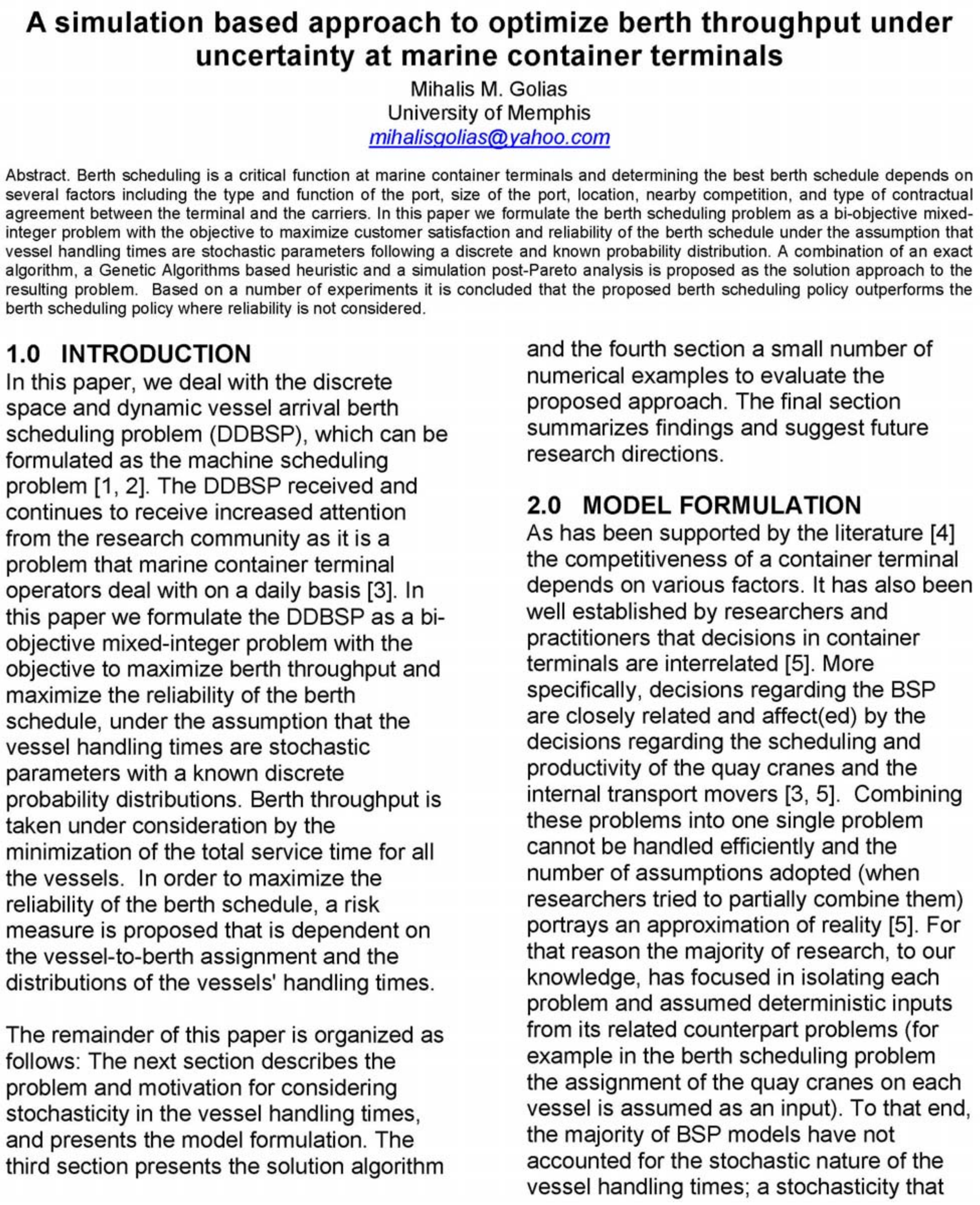
[7] National Business Incubator Association. http://www.nthpower.com/

[8] Moore, Geoffrey A. and Regis McKenna. Crossing the Chasm: Marketing and Selling High-Tech Products to Mainstream Customers. (July 1999).

6.0 ACKNOWLEDGMENT

Dr. Carole Cameron Inge, Chairwoman and CEO of NICCE developed the NICCE business framework and the models described in this paper, along with her colleagues at Virginia Tech University and Research Triangle International. Dr. Inge has been named one of the leading US women in energy and mot recently was the founder of the Virginia Tech Modeling and Simulation Center for Collaborative Technology, the first modeling and simulation center in the university's history. Dr. Inge serves on the US Under Secretary of Energy's Clean Energy Ministerial advisory taskforce called the Clean Energy Education and Empowerment and she is an international speaker and professor for the Practising Law Institute (PLI), a legal education organization chartered by the Regents of the University of the State of New York, founded in 1933. No portion of this paper can be printed, copied or distributed without explicit permission by the author.

2.10 A simulation based approach to optimize berth throughput under uncertainty at marine container terminals



A simulation based approach to optimize berth throughput under uncertainty at marine container terminals

Mihalis M. Golias
University of Memphis
mihalisgolias@yahoo.com

Abstract. Berth scheduling is a critical function at marine container terminals and determining the best berth schedule depends on several factors including the type and function of the port, size of the port, location, nearby competition, and type of contractual agreement between the terminal and the carriers. In this paper we formulate the berth scheduling problem as a bi-objective mixed-integer problem with the objective to maximize customer satisfaction and reliability of the berth schedule under the assumption that vessel handling times are stochastic parameters following a discrete and known probability distribution. A combination of an exact algorithm, a Genetic Algorithms based heuristic and a simulation post-Pareto analysis is proposed as the solution approach to the resulting problem. Based on a number of experiments it is concluded that the proposed berth scheduling policy outperforms the berth scheduling policy where reliability is not considered.

1.0 INTRODUCTION

In this paper, we deal with the discrete space and dynamic vessel arrival berth scheduling problem (DDBSP), which can be formulated as the machine scheduling problem [1, 2]. The DDBSP received and continues to receive increased attention from the research community as it is a problem that marine container terminal operators deal with on a daily basis [3]. In this paper we formulate the DDBSP as a bi-objective mixed-integer problem with the objective to maximize berth throughput and maximize the reliability of the berth schedule, under the assumption that the vessel handling times are stochastic parameters with a known discrete probability distributions. Berth throughput is taken under consideration by the minimization of the total service time for all the vessels. In order to maximize the reliability of the berth schedule, a risk measure is proposed that is dependent on the vessel-to-berth assignment and the distributions of the vessels' handling times.

The remainder of this paper is organized as follows: The next section describes the problem and motivation for considering stochasticity in the vessel handling times, and presents the model formulation. The third section presents the solution algorithm and the fourth section a small number of numerical examples to evaluate the proposed approach. The final section summarizes findings and suggest future research directions.

2.0 MODEL FORMULATION

As has been supported by the literature [4] the competitiveness of a container terminal depends on various factors. It has also been well established by researchers and practitioners that decisions in container terminals are interrelated [5]. More specifically, decisions regarding the BSP are closely related and affect(ed) by the decisions regarding the scheduling and productivity of the quay cranes and the internal transport movers [3, 5]. Combining these problems into one single problem cannot be handled efficiently and the number of assumptions adopted (when researchers tried to partially combine them) portrays an approximation of reality [5]. For that reason the majority of research, to our knowledge, has focused in isolating each problem and assumed deterministic inputs from its related counterpart problems (for example in the berth scheduling problem the assignment of the quay cranes on each vessel is assumed as an input). To that end, the majority of BSP models have not accounted for the stochastic nature of the vessel handling times; a stochasticity that stems from the fact that quay cranes and internal transport vehicles serving the vessels do not have a deterministic productivity (e.g. random down time of quay cranes, unpredicted congestion in the yard, etc). The only exceptions have been three separate studies by Moorthy and Teo [6], Golias et al. [7], and Zhou and Kang [8]. Unlike the model presented herein though, Zhou and Kang [8] proposed a non-linear model formulation minimizing only the total waiting time, Golias et al. [7] focused on online conceptual formulations, and Moorthy and Teo [6] proposed an approach focusing in the randomness of the vessels arrival times and which is relevant only when a substantial number of vessels arrive periodically (as stated by the authors).

In this paper we propose a linear mixed integer bi-objective formulation where we account for the stochasticity in the vessel handling times and assume they are stochastic variables following different discrete probability distributions. The vessel handling time distributions at each berth can be obtained from historical data (i.e. berth assignment, number of QCs and ITVs, breakdown rates of QCs, utilization of yard, vessel handling volumes etc) using data mining algorithms but in this paper we assume that they are known for all the vessels at all the berths. Based on these distributions a risk measure for the berth schedule is proposed and minimized at the same time with the total service time for all the vessels. We choose to introduce the risk measure in contrast to formulating a stochastic optimization problem as the inherent combinatorial complexity of such a model would make it impossible to construct a meaningful heuristic that would efficiently search through the extremely large set of vessel handling time scenarios.

The two objectives introduced, when conflicting, will cause the improvement of one objective to degrade the performance of the other; thus the terminal operator needs to select a schedule that balances between the two objectives. Berth schedules with a high berth throughput have a greater degree of risk (i.e. risk of matching the total service time when the stochastic vessel handling times are realized). On the other hand berth schedules with a lesser degree of risk (decreased berth throughputs), provide more confidence to the terminal operator that the resulting assignment will be stable in terms of the handling times for each vessel and thus deviations from the initial schedule will be minimized in case rescheduling is needed. The proposed model formulation and solution algorithm will provide the terminal operator with the berth schedule that optimally balances between the two objectives. In the following subsection we introduce the risk function and the full model formulation, followed by the solution algorithm in section 3.

2.1 Estimation of berth schedule risk

Let $M_{ij} = \{c_{ij}^1, c_{ij}^2, \ldots c_{ij}^m\}$ be the set of the *m* possible handling times of vessel *j* at berth *i*. Also let $P(c_{ij}^a)$ be the probability that: $c_{ij} = c_{ij}^a \in M_{ij}, \sum_{c_{ij}^a \in M_{ij}} P(c_{ij}^a)$. Then the expected handling time at berth *i* for vessel *j* is equal to: $E(c_{ij}) = \sum_{c_{ij}^a \in M_{ij}} c_{ij}^a P(c_{ij}^a)$. In this paper we define as the measure of risk of a vessel *j* served at berth *i* as: $R_{ij} = \sum_{c_{ij}^a \in M_{ij}} \{\max(0, (c_{ij}^a - E(c_{ij}))P(c_{ij}^a)\}$. We demonstrate this notion by a simple example with one vessel and two berths. Table 1 summarizes the data and results for this example. If the vessel is served at berth 1 then there is a 10% probability that the handling time will be 25 hours, an 80% probability that the handling time will be 31 hours and so on. On the other hand if the vessel is served at berth 2 then there is a 70% probability that the handling time will be 22 hours, a 10% that the handling time will be 25 hours and so on. Although serving the vessel at berth 2 has the lowest expected handling time, it also has the highest risk of exceeding the expected handling time.

Table 1. Example of berth schedule risk estimation for one vessel and two berths

	Berth 1 (*i*=1,*j*=1)			Berth 2 (*i*=2 *j*=1)		
m	c_{ij}^{m}	$P(c_{ij}^{m})$	R_{ij}	c_{ij}^{m}	$P(c_{ij}^{m})$	R_{ij}
1	25	10%	-	22	70%	-
2	31	80%	0.0	25	10%	-
3	35	5%	0.19	38	9%	1.07
4	43	5%	0.59	43	11%	1.86
	$E(c_{ij})=31$		0.78	$E(c_{ij})=26$		2.94

In order to formulate our problem we further define the following:

Nomenclature

Sets

I,J set of berths and vessels

Decision Variables

$x_{ij}\in\{0,1\}, i\in I, j\in J$ 1 if vessel *j* is served at berth *i* and zero otherwise

$y_{ab}\in\{0,1\}, a,b\in J$ 1 if vessel *b* is served at the same berth as vessel *a* as its immediate successor and zero otherwise

$f_j\in\{0,1\}, j\in J$ 1 if vessel *j* is the first vessel to be served at its assigned berth

$l_j\in\{0,1\}, j\in J$ 1 if vessel *j* is the last vessel to be served at its assigned berth

$t_j\in R^{+}, j\in J$ start time of service for vessel *j*

Parameters

$c_{ij}^{a}\in M_{ij}, i\in J, j\in J$ handling time of vessel *j* at berth *i* with probability $P(c_{ij}^{a})$

$A_j, j\in J$ arrival time of vessel *j*

$S_i, i\in I$ time berth *i* becomes available for the first time in the planning horizon

The bi-objective model formulation minimizing the vessel total service time and risk (from now on referred to as RSBM) is formulated as follows:

$$f_1:\min\left[\sum_{j\in J}t_j+\sum_{i\in I}\sum_{j\in J}(E(c_{ij})x_{ij})\right] \qquad (1)$$

$$f_2:\min\sum_{i\in I}\sum_{j\in J}x_{ij}R_{ij} \qquad (2)$$

Subject to:

$$\sum_{i\in I}x_{ij}=1, \forall j\in J \qquad (3)$$

$$f_b+\sum_{a\neq b\in J}y_{ab}=1, \forall b\in J \qquad (4)$$

$$l_a+\sum_{b\neq a\in J}y_{ab}=1, \forall a\in J \qquad (5)$$

$$f_a+f_b\leq 3-x_{ia}-x_{ib}, \forall i\in I, a,b\in J, a\neq b \qquad (6)$$

$$l_a+l_b\leq 3-x_{ia}-x_{ib}, \forall i\in I, a,b\in J, a\neq b \qquad (7)$$

$$y_{ab}-1\leq x_{ia}-x_{ib}\leq 1-y_{ab}, \forall i\in I, a,b\in J, a\neq b \qquad (8)$$

$$t_j\geq A_j, \forall j\in J \qquad (9)$$

$$t_j\geq S_i x_{ij}, \forall i\in I, j\in J \qquad (10)$$

$$t_b\geq t_a+\sum_{i\in I}E(c_{ia})x_{ia}-M(1-y_{ab}), \quad \forall a,b\in J, a\neq b \qquad (11)$$

The first objective function (1) minimizes the expected total service time for all the vessels (from now on referred to as the expected cost or EC). The second objective function (2) minimizes the total risk of exceeding the expected value of handling time for the vessels. Constraint set (3) ensures that each vessel will be served once, while constraint set (4) ensures that each vessel will either be served first or be preceded by another vessel. In a similar manner constraint set (5) ensures that each vessel will either be last or it will be served before another vessel. Constraint sets (6) and (7) ensure that only one vessel can be served first and last at each berth. Constraint set (8) ensures that a vessel can be served after another vessel only if both are served at the same berth. Constraint sets (9) and (10) ensure that the vessel service start time will be greater than the vessel arrival or the time that the berth where the vessel is served becomes available for the first time in the planning horizon. Constraint set (11) estimates the start time of service for each vessel.

3.0 SOLUTION ALGORITHM

The RSBM is a bi-objective optimization problem and for both single objective problems (derived once we eliminate one of the two objectives) no exact solution algorithm exist that can be applied and solve them in polynomial time. In order to tackle this issue a new heuristic approach is presented. The proposed heuristic is an improvement of the exact algorithm 2-PPM proposed by Lemerse et al., [9]. The concept of the 2-PPM algorithm was to split the search space into equal and predetermined partitions and then for each partition use the *ε*-constraint method to find a solution. The algorithm proposed herein follows the same concept of partitioning the search space, but does so in an adaptive manner without having to predefine the size or the number of partitions. Furthermore, instead of the *ε*-constraint method the weighted approach is used to produce a solution within each partition, resulting in a faster estimation of the PF (PF). Before we present the proposed heuristic we define the following:

Definition 1: Let *X* is the feasible space of the RSBM and $x \in X$ be a feasible solution. Solution $a \in X$ dominates solution $b \in X$ if: $\{f_1(a) \le f_1(b), f_2(a) < f_2(b)\}$ or $\{f_1(a) < f_1(b), f_2(a) \le f_2(b)\}$. Any dominated solutions do not belong in the PF (the set of PF solutions is denoted from now on as Ω).

Definition 3: Let $x_n^{pf} \in \Omega \subset X$ be the *n*th PF solution of the RSBM.

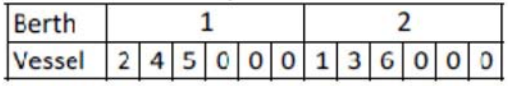
The proposed heuristic that can provide Ω (from now on called the Bi-objective Berth Scheduling Heuristic or BBSH), can be described using the pseudo-code in figure 1. As the single objective problems (solved at step 2 of the BBSH) are *NP*-Hard, a Genetic Algorithms (GAs) based heuristic, proposed for the DDBSP by Golias et al., [10] is employed as the solution algorithm. The GAs heuristic is briefly described in the following subsection for consistency purposes.

```
Step 1: Set $Z_{new} = \varnothing$
$\Omega = Z = (x_1^{pf} : \arg\min_{x \in X} f_1(x), x_2^{pf} : \arg\min_{x \in X} f_2(x))$
$\Pi_1 = f_1(x_1^{pf}), \Pi_2 = f_2(x_2^{pf})$
Step 2:
for i=2: |Z|
        if $\left| \frac{f_1(x_i^{pf}) - f_1(x_{i-1}^{pf})}{f_1(x_i^{pf})} \right| > 0.01$ and
        $\left| \frac{f_2(x_{i-1}^{pf}) - f_2(x_i^{pf})}{f_2(x_{i-1}^{pf})} \right| > 0.01$
        $P : x_{|Z|+1}^{pf} = \arg\min_{x \in X} \left( \frac{f_1(x)}{\Pi_1} + \frac{f_2(x)}{\Pi_2} \right)$
        $s.t.$
        $f_1(x_{i-1}^{pf}) < f_1(x) < f_1(x_i^{pf})$
        $f_2(x_i^{pf}) < f_2(x) < f_2(x_{i-1}^{pf})$
                if P is infeasible: $Z = Z \setminus x_{|Z|-1}^{pf}$
                else
                $Z_{new} = Z_{new} \cup x_{|Z|+1}^{pf}, \Omega = \Omega \cup x_{|Z|+1}^{pf}$
                end if
        end if
end for
$Z = Z \cup Z_{new}, Z_{new} = \varnothing$
Order and renumber solutions in Z based on tuple
$\{f_1(x), f_2(x)\}, x \in Z$
Step 3: If |Z|>1 go to step 2 else end
```

Figure 1. BBSH Pseudo-code

3.1 GAs heuristic

The GAs heuristic consists of four parts: a) the chromosomal representation, b) the chromosomal mutation, c) the fitness evaluation and d) the selection process. For scheduling problems integer chromosomal representation is more adequate, since the classical binary representation can obscure the nature of the search [11]. In this paper, we use an integer chromosomal representation, in order to exploit in full the characteristics of the problem. An illustration of the chromosome structure is given in figure 2 for a small instance of the problem with 6 vessels and 2 berths. As seen in figure 2 the chromosome has twelve cells. The first 6 cells represent the 6 possible service orders at berth1 and the last 6 cells the 6 possible service orders at berth 2. In the assignment illustrated in figure 2, vessels 2, 4, and 5 are served at berth 1 as the first, second and third vessel respectively, while vessels 1, 3, and 6 are served at berth 2 as the first, second, and third vessel respectively.



Berth	1						2					
Vessel	2	4	5	0	0	0	1	3	6	0	0	0

Figure 2. Illustration of chromosome representation

Four different types of mutation are applied as part of the genetic operations to all the chromosomes at each generation: insert, swap, inversion, and scramble mutations. Since the RSBM is a minimization problem the smaller the values of each objective function are, the higher the fitness value will be. We define the fitness function of a chromosome as: $f_{pt}(x) = F^{pt}_{\max} - z_{pt}(x)$, where $F^{pt}_{\max}$ is the maximum value of the objective function z_{it}, and f_{it} is the value of the fitness function of objective function *i* at iteration *t* of the GA. However, this value is not know in advance and is replaced by the largest value z_{pt} of each objective function at each iteration. Chromosomes whose objective function values do not satisfy the constraints of problem *P* are replaced by their parents.

3.2 Post Pareto simulation

The algorithm described in the previous subsections will produce a number of non-dominated solutions (i.e. berth schedules of the PF). The next step will be to select one of these solutions as the schedule to be implemented. This follow up step is known as post-Pareto analysis and can be quite a challenging task since, in the absence of subjective or judgmental information, none of the corresponding trade-offs can be said to be better than the others [13]. In the problem studied herein we employ simulation as means to select one schedule from the PF that will be implemented. The simulation entails the use of a simple Monte Carlo procedure that generates random instances of the vessels handling times and estimates an average of the total service time over all the instances. The procedure can be described as follows. Let *K* be the total number of different handling time instances (i.e. realizations of the vessels handling time) we wish to produce and $CPF_{ij}(\cdot)$ the cumulative distribution function of vessels' *j* handling time at berth *i*. This procedure is shown in figure 3.

```
for $k$=1:$K$, $i$=1:$|I|$, $j$=1:$|J|$
        Generate a number from the
        Uniform Distribution [0,1]: $u=U(0,1)$
        Set $cs^k_{ij} = CPF_{ij}(u)$ , where: $cs^k_{ij}$ : is
        the $k^{th}$ realization of vessels' $j$
        handling time at berth $i$
end
```

Figure 3. Monte Carlo Procedure (MCP)

For each one of the solutions in the PF we estimate the mean value of the total service time (from now on referred to as the mean simulated cost or *MSC*) over all the *K* realizations of the vessel handling times (obtained from the MCP) as:

$$MSC(x^{pf}_n) = \sum_{k=1}^{K} f_1(cs^k_{ij}, x^{pf}_n) \Big/ K$$

. The solutions with the minimum *MSC* over all the schedules in the PF (from now on referred to as the Pruned PF Solution or PPFS and denoted by x^{ppfs}) is selected as the schedule to be implemented.

4.0 COMPUTATIONAL EXAMPLES

Problems used in the experiments were generated randomly, but systematically. We developed twenty base problem instances, where vessels are served with various handling volumes at a multi-user container terminal (MUT) with five berths, with a planning horizon of one week, and various handling volumes. The range of variables and parameters considered herein were chosen according to [10, 13]. As previously discussed in subsection 2.1 of this paper we assume that vessel handling times are stochastic parameters following a discrete probability distribution. Without loss of generality for the computational examples presented herein we assumed that each vessel at each berth will have a maximum of five possible handling times (i.e. *m=5*). We assumed that the handling time at each berth for each vessel increases randomly (from the minimum handling time) based on a uniform probability distribution with a minimum of 5% and a maximum of 15% (i.e.: if c_{ij}^1 is the minimum handling time of vessel *j* at berth *i* then the second through the fifth other possible handling times are estimated as: $c_{ij}^2 = c_{ij}^1 \times U(1.05,1.15)$, $c_{ij}^3 = c_{ij}^2 \times U(1.05,1.15)$, $c_{ij}^4 = c_{ij}^3 \times U(1.05,1.15)$, $c_{ij}^5 = c_{ij}^4 \times U(1.05,1.15)$). In total 5 datasets where created for each of the two different vessel inter-arrival times with different handling volumes.

For each dataset, and out of these five possible handling times at each berth, we randomly assigned the handling time that will receive the highest probability (from now on referred to as the dominant handling time or c_{ij}^d and dominant probability or $P(c_{ij}^d)$ respectively). For each dataset we developed four different cases where we allowed $P(c_{ij}^d)$ to vary based on the values shown in table 2. The probabilities for the four remaining handling times were estimated using the procedure shown in figure 4.

Table 2. Dominant handling time probabilities

Dominant Probability	50%-60%	60%-70%	70%-80%	80%-90%

In total 50 problem instances were created. Concerning the GAs based heuristic the population size was set to 50 chromosomes and the GA-based heuristic is stopped if no improvement is observed for 100 iterations (i.e. new or improved schedules found). Finally, the algorithm was implemented in Matlab and the experiments were performed on an ASUS desktop personal computer (E5300@2.60GHz) with 6GB memory.

$$Step\_1: P(c_{ij}^m) = U(P(c_{ij}^d),1), \forall c_{ij}^m \neq c_{ij}^d \in M_{ij}$$

$$Step\_2: P(c_{ij}^m) = \frac{1-P(c_{ij}^d)}{\sum_{c_{ij}^m \neq c_{ij}^d \in M_{ij}} P(c_{ij}^m)} P(c_{ij}^m)$$

Figure 4. Handling Time Probabilities Estimation Procedure

4.1 Evaluation of berth scheduling policy

In this subsection we evaluate the payoff of introducing the second objective function. For each one of the 50 problem instances, previously described, we obtained the PF using the heuristic algorithm presented in section 3. For each schedule in the PF we calculated the *MSC* over a sample size of *K*=10 000. As discussed in section 3.2 the schedule to be implemented will be the one with the minimum *MSC*. To evaluate the effectiveness of the proposed policy we compared the $MSC(x^{ppfs})$ value to the *MSC* value of the solution in the PF with the minimum EC. The latter is the solution: $x_1^{pf} : \underset{x \in X}{\operatorname{argmin}} f_1(x)$ (i.e. the solution we would obtain if we did not consider the risk function as the problem would be single objective) and from now on will be referred to as the Nadir PF Schedule or NPFS and denoted by x^{npfs}. Table 3 reports the results from the comparison between $MSC(x^{ppfs})$ and $MSC(x^{npfs})$ values for all the 200 problem instances. The percentages reported in table 3 are estimated as: $\frac{MSC(x^{npfs}) - MSC(x^{ppfs})}{MSC(x^{npfs})}$, and answer the following question: "*On average should we expect a gain in berth throughput if we choose the PPFS over the NPFS and by how much?*". From the results in table 3 we observe that the PPFS always produces a smaller *MSC*. For example for the first dataset and the first dominant probability (i.e. 50%-60%) the *MSC* of the NPFS solution is 12% and 3% larger than the *MSC* of the PPFS solution for the 3 and 5 hours vessel inter-arrival times respectively.

Table 3. *MSC* values difference (in %) between the NPFSs and the PPFSs (3 and 5 hours of vessel inter-arrival)

Dataset	50%-60%	60%-70%	70%-80%	80%-90%
1	12 / 3	19 / 27	9 / 13	12 / 31
2	15 / 31	10 / 49	30 / 57	4 / 34
3	10 / 44	2 / 41	34 / 18	9 / 12
4	44 / 32	7 / 33	12 / 20	8 / 9
5	24 / 14	37 / 0	42 / 29	7 / 63

5.0 CONCLUSIONS

In this paper, we formulated the discrete space and dynamic vessel arrival berth scheduling problem as a bi-objective mixed-integer problem with the objective to maximize the berth throughput and the reliability of the berth schedule, under the assumption that vessel handling times are stochastic parameters with known discrete probability distributions. In order to maximize the reliability of the berth schedule, a measure of risk was proposed dependent on the vessel-to-berth assignment. In order to solve the resulting problem, a combination of an exact and a GAs based heuristic were proposed and a number of simulation experiments were performed. Based on results from these experiments it was concluded that considering risk (from the inherent stochasticity of the vessel handling times) in berth scheduling can provide schedules with improved berth throughput when the vessel handling times are realized. Future research is focusing in: a) in a model formulation where the mutual impact of the vessels' stochastic handling times are considered, and b) evaluation of the proposed framework in terms of the robustness, dominance, and expected loss of the final schedule over all the Pareto points.

6.0 REFERENCES

[1] Imai, A., Nagaiwa, K., Tat, C-W., 1997. Efficient planning of berth allocation for container terminals in Asia. Journal of Advanced Transportation 31, 75–94.

[2] Imai, A., Nishimura, E., Papadimitriou S., 2001. The dynamic berth allocation problem for a container port. Transportation Research Part B, 35,401–417.

[3] Meisel, F., 2009. Seaside operations planning in container terminals. Physica-Verlag Berlin Heidelberg.

[4] Bierwirth, C., Meisel, F., 2009. A survey of berth allocation and quay crane scheduling problems in container terminals. European Journal of Operational Research, 202(3), 615-627.

[5] Ballis, A., Dimitriou, L., Paravantis, J., 2010. Quay to Storage Area Container Transfer: Critical Review of Modeling Techniques and Practical Outcomes. In: Proceedings of the 89th Annual Meeting of the Transportation Research Board, 10-3020.

[6] Moorthy, R., Teo C-P., 2006. Berth management in container terminal: The template design problem. OR Spectrum 28(4), 495-518.

[7] Golias, M.M., Boilé, M., Theofanis, S., 2007. The stochastic berth scheduling problem. In: Proceedings of the Second TRANSTEC Conference, Prague, Czech Republic.

[8] Zhou P-F, Kang, H-G., 2008. Study on berth and quay crane allocation under stochastic environments in container terminal. Systems Engineering Theory and Practice 28(1),161-169.

[9] Lemesre, J., Dhaenens, C., Talbi, E.G., 2007. Parallel partitioning method (PPM): A new exact method to solve bi-objective problems. Computers and Operations Research 34,2450–2462.

[10] Golias, M.M., Boilé, M., Theofanis, S., 2009. Service time based customer differentiation berth scheduling. Transportation Research Part E 45(6),878-892.

[11] Eiben, A.E., Smith, J.E., 2003. Introduction to Evolutionary Computing. Springer.

[12] Golias, M.M., Boilé, M., Theofanis, S., Taboada, A.H., 2010. A multi-objective decision and analysis approach for the berth scheduling problem. International Journal of Information Technology Project Management 1(1), 54-73.

[13] Hansen, P., Oguz, C., Mladenovic, N., 2008. Variable neighborhood search for minimum cost berth allocation. European Journal of Operational Research 131(3), 636-649.

7.0 ACKNOWLEDGMENT(S)

This material has been partially supported by the Intermodal Freight Transportation Institute, at the University of Memphis. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Intermodal Freight Transportation Institute.

2.11 Reducing US Oil Dependence Using Simulation

Reducing US Oil Dependence Using Simulation

Fadi Ayoub - Ph.D. Student
Old Dominion University
fayou001@odu.edu

Georges M. Arnaout - Ph.D. Candidate
Old Dominion University
garna001@odu.edu

Abstract **.**People across the world are addicted to oil; as a result, the instability of oil prices and the shortage of oil reserves have influenced human behaviors and global businesses. Today, the United States makes up only 5% of the global population, but consumes 25% of the world total energy. Most of this energy is generated from fossil fuels in the form of electricity.
The contribution of this paper is to examine the possibilities of replacing fossil fuel with renewable energies to generate electricity as well as to examine other methods to reduce oil and gas consumption. We propose a system dynamics model in an attempt to predict the future US dependence on fossil fuels by using renewable energy resources such as, nuclear, Wind, solar, and hydro powers. Based on the findings of our model, the study expects to provide insights towards promising solutions of the oil dependency problem.

1.0 INTRODUCTION

Oil shortage is affecting various facets of our lives, such as the economy, the environment, national security, government policies and human behaviors. According to the Energy Information Administration report 2009, U.S. fossil fuel consumption decreased by 4.2%; however, the major petroleum product especially gasoline did not decline, in fact, the average annual consumption is predicted to increase by 20,000 bbl/d and 90,000 bbl/d in 2010 and 2011, respectively [1]. The report implies that oil will continually be produced and imported at increasing rates in order to meet the rising consumption levels while oil reserves are limited. As a result, the health of our future will ultimately depend on all available energy resources to include oil, and renewable energies.

Investing in alternative energy sources has become popular in the recent decade. The main reason behind such investments is to decrease American oil dependency on imported foreign oil by generating substitutable energies for power generation from different avenues, such as from wind turbines and natural gas. In this paper, we propose a system dynamics model that represents the US energy consumption, generation, and use of alternative energy sources focusing mostly on electricity generation from alternative renewable sources. The aim of this paper is to examine the future US oil dependence by using renewable energy sources. The study expects to provide insights towards promising solutions of the US oil dependency problem.

2.0 PROBLEM STATEMENT

Recently, researchers have been focusing on creating and generating alternative renewable energies which can reduce the amount of imported foreign oil. This could provide advantages in term of employment rate, environmental sustainability, technological development and local economies; however, such strategies require a huge investment and support from both the private sector as well as the government. In this study, we propose a simplistic model of energy generation and consumption in the US focusing on electricity generation from alternative renewable energy sources. In order to examine the model, the data collection have been retrieved from the Energy Information Administration Independent Statistics and Analysis. The influenced factors include oil production and consumption, gas production and consumption, electricity consumption and generation, oil consumption to electricity generation, alternative energies consumption and production (hydropower, nuclear, wind, thermo, solar), electric car and light bulb (CFL). The data will be analyzed quantitatively and conclusions will be drawn according to the model's results.

3.0 MODEL CAUSAL LOOP

One of the intentions of the proposed model is to predict the behavior of the transition from traditional non-renewable to renewable sources of energy. In consequence, we take into consideration the three main non-renewable sources used nowadays: oil, gas and coal. As it has been evident through the years, the demand for energy is constantly increasing. In our model, we consider the demand from four main sectors: transportation, commercial, industrial, and residential. The high increase in the demand for energy and the decrease, since many years ago, in the petroleum extraction have alerted the world about the critical need to find and use alternative energy sources. In this model, we consider the main renewable sources that are starting to emerge and become important in electricity generation for some countries; including The United States. The renewable sources considered in this paper are: wind, nuclear, hydro, thermo and solar sources. Furthermore, another way to preserve and optimize the use of petroleum is by utilizing more efficient technologies that use less energy and provide the same benefits. Therefore, in this model we consider the impact of using technologies such as: LED's and electric cars, which help in reducing the amount of energy that is currently obtained from fossil sources.

A simple way to grasp our model is by reviewing the casual loop diagram (shown in fig.1) which; while avoiding to mathematically validate our predictions, is a summary of the factors considered that will have an impact on the consumption, usage, and availability of fossil and renewable resources. According to the casual loop diagram, when the residential, commercial or industrial demand increases, the oil, gas and coal consumption also increases; therefore, the relation is positive. In the same way, if the oil, gas and/or coal consumption increases, then the oil, gas and/or coal resources decrease, respectively; in consequence, the relation is negative. Following the diagram, it is evident that if any of these resources decline; the non-renewable resources reservoir would decrease as well, implying a positive relation. Retrospectively, increase in consumption from alternative sources would yield a decrease in the non-renewable sources consumption (negative relation). Following the same logic, having more availability of alternate resources would allow us to have a higher electricity generation from renewable sources (positive relation) and reduce the consumption of fossil sources to generate electricity (negative relation). Finally, an increase in the use of saving light bulbs will also help to decrease the electricity generation from fossils (negative relation) in the same way that using more electric cars will allow us to reduce the consumption of oil for transportation (negative relation).

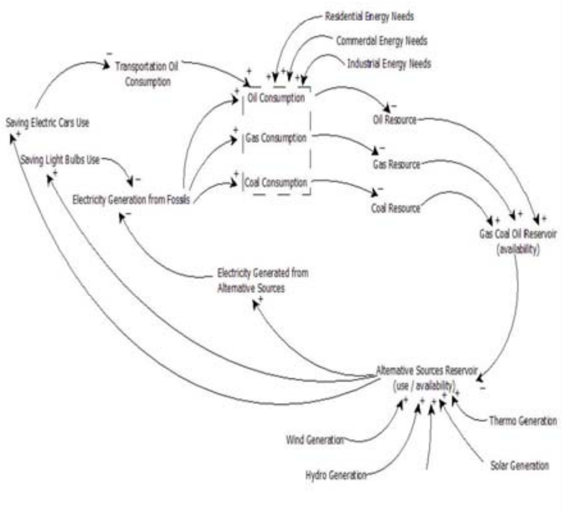


Figure 1. Model's Causal Loop diagram

The proposed model is created using GoldSim® version 10.10, a Monte Carlo simulation software based on system dynamic modeling [3]. In this section, we explain the technical details of the model we developed. The section is divided between Model Limitations and Model Technical Details.

4.1 Model Limitations

The model assumptions and limitations are presented below:

a- The model is constrained within the time frame 1980-2035.

b- High-Efficiency Light Bulbs were considered to yield 20% saving in electricity consumption when compared to standard light bulbs.

c- Each electric vehicle was assumed to save an average of 600 gallons annually based on an average of 12000 travel miles per year with a consumption of 20 miles per gallon.

d- Renewable resources were assumed to be installed upon demand without any limitations.

e- Only 10% of the available area of the wind maps was considered for wind turbine installations.

f- Solar panels were modeled without any limitations, such as size, locations, or available times.

g- Increase in electric consumption due to electric vehicles was considered.

h- Geothermal heat was modeled without any limitations.

i- All modeling and simulation for future purposes was based on the year 2007 data.

j- While the data in GoldSim is shown in bbl, it is actually in 1000 bbl. We did this for simplification of the graph readings.

k- Nuclear plants take on average 10 years between being built and being commissioned, in order to start producing energy. Although we haven't modeled this important constraint, we adopted a lower percentage for nuclear reactors future production taking this delay factor into consideration.

l- Although in real life the oil and gas reserves are rarely altered, we assumed that our reserves are in a tank (or reservoir to be more accurate) where the consumption is withdrawn from it and the import is added to it.

m- In our model, we are assuming that the increase is constant (e.g. the same percentage increase of wind turbines and solar panels are added every year on the previous one).

n- The future alternative energy sources are assumed to start at year 2007 and the data accuracy of the prediction is dependant on [4].

o- The future increase of the energy produced by alternative energy sources starts from year 2007 and until 2035.

p- Due to lack of data, future prediction for electric cars was based on the number of electric cars available in 2007 [4] consuming 150 watts/hr per car for total of 5 hours per day. In addition, we assume that based on the predicted rate increase in electric cars in the market, a similar decrease would be witnessed in fossil-fuel based vehicles. The fossil-fuel vehicles were assumed to consume an average of one gallon per 20 miles for a total of 12000 miles driven annually per car.

4.2 Model Technical Details

The model is designed to reflect both the current and the future state of the energy production and consumption sectors. The current state represents the *no change* in culture "status quo" scenario, while the future state represents the implementation of alternative energy resources to include high efficiency systems. Below in fig. 2, is a layout that shows the blue print of our design.

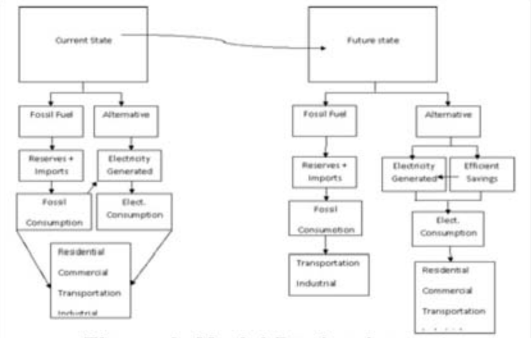


Figure 2. Model Design Layout

The initial high-level view of the model looks like fig. 3. The model is divided into four main containers and one control panel allowing the modeler to control the sliders of the future increases in alternative energy sources.

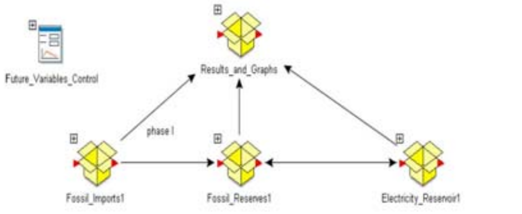


Figure 3. Goldsim Model

5.0 RESULTS AND ANALYSIS

When considering a constant increase approach for the future alternative energy sources, we compared the future electricity generation from alternative sources with the predicted electricity consumption. After increasing the future energy sources by certain percentages in the control panel (shown in fig. 4), we ran the system and compared the alternative electricity generation with the predicted electricity consumption. Note that we put the maximum amount for the variables in the control panel that according to our belief, are feasible.

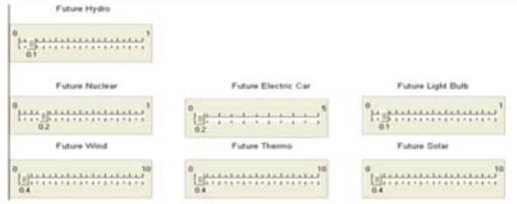


Figure 4 Control panel

When looking at the results' graph (shown in fig. 5), we need to take into consideration that the future alternative sources start from year 2007 (i.e. reference year 0 on the graph) until the year 2035 (i.e. year 28 on the graph) meaning that at year 2035 for instance, we will generate 75130280 * 1000 bbl.

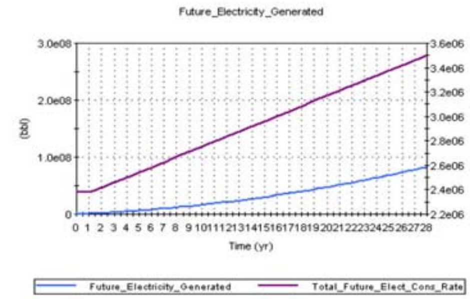


Figure 5. Comparison Graph

As observed, according to our study, when considering these 3 alternatives, i.e. wind, solar, and geothermal resources, it is not feasible to achieve, our objective and meeting the consumption demand. However, when Natural Gas was considered such as a simplistic optimization approach yields that it is a possibility to achieve our goal. Table 1 shows the parameters used to achieve the objective function and meet the energy consumption demand.

Table 1. Parameters Used to Achieve Objective Function

Resource	Preferential weight factor per sector	% increase in annual production per sector
	w	*a/sector*
Wind	0.025	0.0625
Solar Panel	0.025	0.0625
Geothermal	0.025	0.0625
Nuclear	0.775	0.25
Hydro	0.025	0.16
Natural Gas	0.125	0.25

5.1 Optimization

In this section, a Non Linear Programming (NLP) Model was developed in Excel to quickly study the possible strategies for meeting future demand. For this reason we introduced two preferentiality factors i.e., strategy-variables, denoted as ω and α. ω was used to limit the production of a given sector under a given percentage of the total output, where as α was used to limit the annual increase in production of a given sector to a predetermined percentage. For the purpose of this paper we limited α to less than 0.25 while we allowed Excel to determine ω. It is understood that our results are strict to these conditions and that any change in the preferentiality factors would ultimately change them. Equations below were used to establish our optimization model.

$$Z = \sum_{j=1}^{j=6} P_j,$$

$$P_j = X_j + X_j * \partial_j * t_j$$

$\mathit{where}\ \mathrm{P_j} = \mathit{Power\ output}$ per Sector $\mathrm{X_j}$

$$\partial = \text{Production increase rate per anuuam} = \frac{\mathrm{P_{t+1}} - \mathrm{P_t}}{\mathrm{P_t}}$$

Subject to the constraint of $\mathrm{P_j} * \omega_j$ for any sector j

where ω_{j} is a prefentiality factor per sector j

The cost associated with installing additional units (not previously installed) can be calculated as such:

$$C_j = (\sum_{u=1}^{u=k} c_u) * (1+i)^t$$

$$X_j = \sum_{u=1}^{u=k} x_u$$

$\mathit{where}\ c_u =$ cost per a single unit, x_u, of sector J

$i =$ inflation rate or interest rate per year

t = time to reach power output required

subject to the following constraints :

$$\sum_{j=1}^{6} \mathrm{X_j} P_j \geq \text{Predicted Energy Demand in 2035}$$

$$\mathrm{P_j} = \text{Energy Production of sector j} = \sum_{u=1}^{u=k} \mathrm{p}_u x_u$$

$$0 \leq \mathrm{t} \leq 15$$

$$0 \leq \alpha \leq 25\%$$

$$\sum_{j=1}^{6} \omega_{\mathrm{j}} = 1$$

$$0 \leq \omega_j \leq 1$$

$$x_u \geq 1$$

$$1 \leq \mathrm{P_s} \leq 0.1 * \mathit{Current}\ \text{No.}\ \mathit{Homes} * 4\text{KWh}$$

$$1 \leq X_{\mathrm{w}} \leq \frac{0.1 * \mathit{Total}\ \text{wind Area available}}{\text{Area of } \mathrm{x_w}}$$

$$1 \leq X_H \leq 100 + (\text{Currently installed})$$

$$1 \leq X_N \leq 50 + (\text{Currently installed})$$

$$1 \leq X_G$$

$$\text{Currently installed} \leq X_{NG}$$

where w : wind Power Sector

H : Hydro Power Sector

S : Solar Power Sector

N : Nuclear Power Sector

G : Georthermal

and NG : Natural Gas

Some of the assumptions that were taken into accounts for building the model were:

- 10% of the total available wind land would be the maximum area to utilize.
- 10% of the residential homes would be assumed to install solar panels.
- Policies would not restrict or prohibit building of damns or solar panels or wind mills.
- There is assumed to be available resources to complete installation of any selected sector at any quantity.

- Environmental and other policy decisions could be represented by the preferentiality factor.

Other factors that were not considered:

- Operation costs of a plant
- Life cycle of a plant
- Decommissioning cost of a plant

What we now have, in fact, are two optimization objective functions that can be manipulated in the desired perspective. For instance, if cost is the main goal, which is not in our case, then minimizing the cost objective function would yield the sought after results, else our objective function Z is the answer.

One thing that must be kept in mind that all energy units have been represented on the basis of electricity equivalent. Since wind turbines, solar power, nuclear plants, and the remaining sector capacities are valued base on Kwh capacity. (The required 3.5E9 barrels/year in figure 5 was converted to MW). Input and constraints requirements and results for NLP model are given in tables 2 and 3 respectively.

Table 2: Model Inputs and Constraints for Energy Optimization Non-Linear Program

Source	Kwh/Unit/Yr	$ Cost/Unit	Area required ft^2	Size/Numbers Constraints
Wind	3*10^3	10^6	84497	1.73052*10^12 ft^2
Residential Solar	1.4*10^3	36000	120/house	12.6*10^6 houses
Geothermal	1.63*10^8	8.4*10^7	N/A	N/A
Nuclear	12.4 10^9	1.4*10^10	N/A	124
Small Hydro-electric	2.6*10^6	2*10^6	N/A	4139
Natural Gas	1.3*10^6	1.5*10^8	N/A	727

Table 3: Results of the Optimization model

Power Source	# of units	# units installed/already installed in 1st yr	Months to achieve goal	Total production MW	Total cost $
Wind	1	1	5.63	6.00E09	2.96E13
Solar	1	1	5.62		
Geothermal	2	2	5.62		
Nuclear	330	124	79.7		
Hydro	4460	4139	5.8		
Natural Gas	1463	727	48.6		

The NLP optimization Model as demonstrated above shows that it is possible to meet the future demand of energy equivalent to 6E9 MW at a rough estimated cost of 2.962E13 over a period of 4 years.

6.0 CONCLUSION AND DISCUSSION

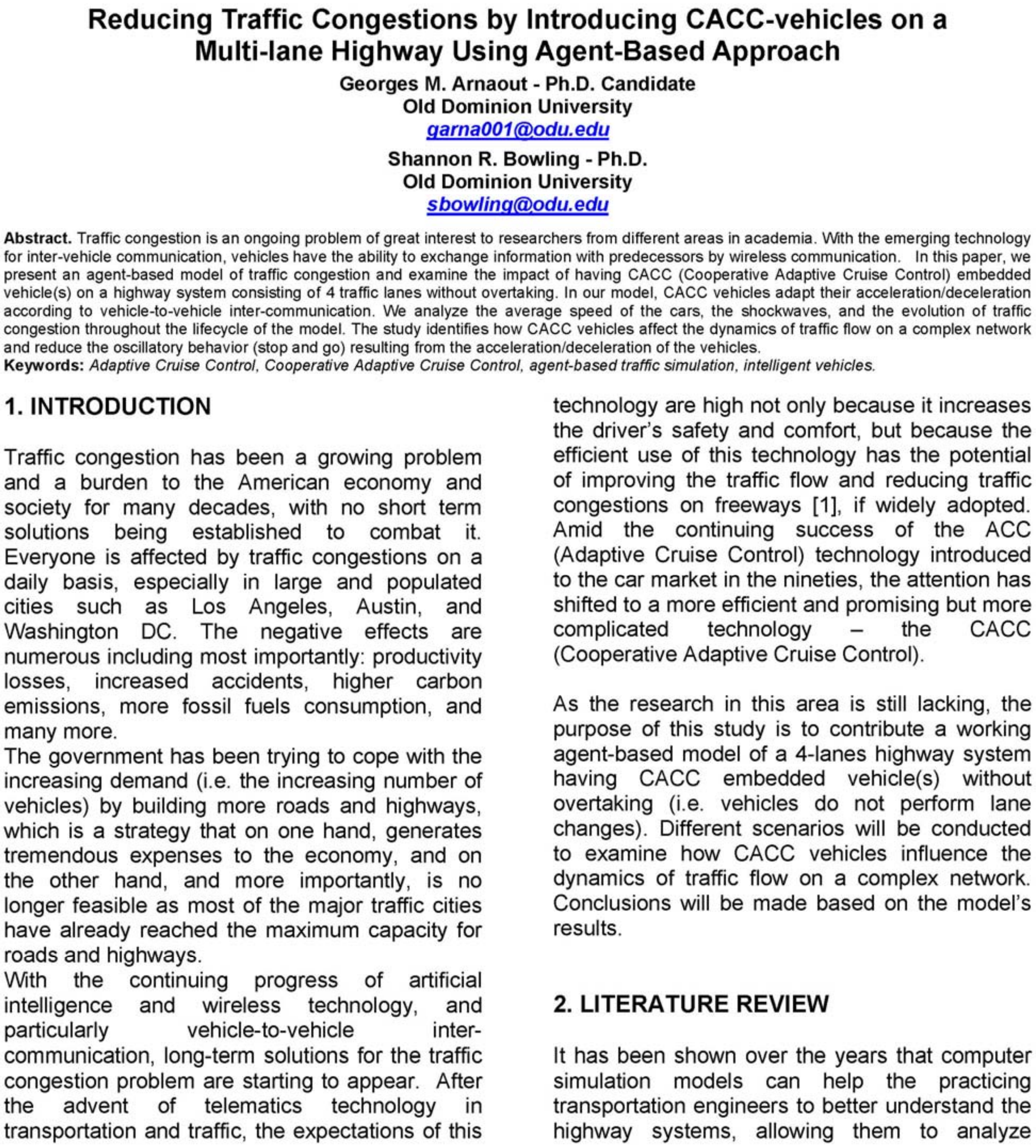
The purpose of this paper was to examine the future US oil dependence by using renewable alternative energy sources focusing on electricity. A simulation model was developed in which we considered wind, solar, and geothermal heat, as the only renewable resources available to generate electricity. Furthermore, we tested the effect of replacing light bulbs with high efficiency light bulb with respect to electricity consumption. Also, instead of using natural gas as the fuel of choice for transportation, electric vehicle technologies were considered for vehicles. No change in fuel consumption was considered for other modes of transportation, such as airlines, trains, large shipping trucks and maritime shipping.

After Setting up the modeling environment with the necessary functions and data, the renewable resources production rates were manipulated. The study concludes that it is possible to be energy independent from foreign oil but at an astronomical costs. A forcing policy must be implemented to entice *all* the private energy companies for seeking alternative energy sources.

REFERENCE

1. Administration, E.I., Energy Information Administration: Official Energy Statistics from the U.S. Government 2009, in Short Term Energy Outlook. 12 March, 2010.
2. Vuong, A., Tycoon's plan taps wind, in The Denver Post. 12 March 2010.
3. Associates, G. GoldSim. June 25, 2010]; Available from: http://www.goldsim.com/.
4. Administration, E.I., Forecasting and Analysis. 2010.

2.12 Reducing Traffic Congestions by Introducing CACC-vehicles on a Multi-lane Highway Using Agent-Based Approach



Reducing Traffic Congestions by Introducing CACC-vehicles on a Multi-lane Highway Using Agent-Based Approach

Georges M. Arnaout - Ph.D. Candidate
Old Dominion University
garna001@odu.edu

Shannon R. Bowling - Ph.D.
Old Dominion University
sbowling@odu.edu

Abstract. Traffic congestion is an ongoing problem of great interest to researchers from different areas in academia. With the emerging technology for inter-vehicle communication, vehicles have the ability to exchange information with predecessors by wireless communication. In this paper, we present an agent-based model of traffic congestion and examine the impact of having CACC (Cooperative Adaptive Cruise Control) embedded vehicle(s) on a highway system consisting of 4 traffic lanes without overtaking. In our model, CACC vehicles adapt their acceleration/deceleration according to vehicle-to-vehicle inter-communication. We analyze the average speed of the cars, the shockwaves, and the evolution of traffic congestion throughout the lifecycle of the model. The study identifies how CACC vehicles affect the dynamics of traffic flow on a complex network and reduce the oscillatory behavior (stop and go) resulting from the acceleration/deceleration of the vehicles.

Keywords: *Adaptive Cruise Control, Cooperative Adaptive Cruise Control, agent-based traffic simulation, intelligent vehicles.*

1. INTRODUCTION

Traffic congestion has been a growing problem and a burden to the American economy and society for many decades, with no short term solutions being established to combat it. Everyone is affected by traffic congestions on a daily basis, especially in large and populated cities such as Los Angeles, Austin, and Washington DC. The negative effects are numerous including most importantly: productivity losses, increased accidents, higher carbon emissions, more fossil fuels consumption, and many more.

The government has been trying to cope with the increasing demand (i.e. the increasing number of vehicles) by building more roads and highways, which is a strategy that on one hand, generates tremendous expenses to the economy, and on the other hand, and more importantly, is no longer feasible as most of the major traffic cities have already reached the maximum capacity for roads and highways.

With the continuing progress of artificial intelligence and wireless technology, and particularly vehicle-to-vehicle inter-communication, long-term solutions for the traffic congestion problem are starting to appear. After the advent of telematics technology in transportation and traffic, the expectations of this technology are high not only because it increases the driver's safety and comfort, but because the efficient use of this technology has the potential of improving the traffic flow and reducing traffic congestions on freeways [1], if widely adopted. Amid the continuing success of the ACC (Adaptive Cruise Control) technology introduced to the car market in the nineties, the attention has shifted to a more efficient and promising but more complicated technology – the CACC (Cooperative Adaptive Cruise Control).

As the research in this area is still lacking, the purpose of this study is to contribute a working agent-based model of a 4-lanes highway system having CACC embedded vehicle(s) without overtaking (i.e. vehicles do not perform lane changes). Different scenarios will be conducted to examine how CACC vehicles influence the dynamics of traffic flow on a complex network. Conclusions will be made based on the model's results.

2. LITERATURE REVIEW

It has been shown over the years that computer simulation models can help the practicing transportation engineers to better understand the highway systems, allowing them to analyze everyday traffic management needs by looking at congestion problems and understanding their cause. Traffic simulation allows transportation engineers to examine behaviors that are not apparent through visual observation. Three types of simulations have been found in the literature pertaining to traffic simulation: (1) Microscopic – model focuses on individual vehicle progress and movements; (2) Macroscopic – model focuses on the overall traffic flow such as the mean speed, variance, flow rate, density, etc ; and (3) Mesoscopic – model focuses on individual vehicles at the aggregate level by speed density relationship and queuing theory approaches [2]. According to the literature [3] and [4] , ACC could have both, a positive as well as a negative effect on the traffic flow. [5] examine how a low penetration level of ACC does not have any effect on traffic flow, regardless of the time-gap set.. With a relatively high penetration of ACC (between 20% and 60%) in the market, and even under the most advantageous conditions, an ideal ACC system can only have a small impact on highway capacity [6]. ACC vehicles are currently developed by the car industry and their penetration is relatively low. With an average time gap of 1.4 seconds between ACC and other vehicles, the highway capacity can be increased by at most 7% [6]. Additionally, the same study has shown that with a penetration of ACC above 60%, the effect of ACC can be negative and can lead to a modest loss of highway capacity.

Unlike ACC, the literature pertaining to CACC is limited. Several studies examined CACC designs and architectures but most of the studies did not explore the traffic flow effects of CACC in terms of throughput, capacity, and congestion reduction. One of the very few studies that targeted this area of research was [6], where it identified that CACC vehicles enable closer vehicle following (time gap as low as 0.5 seconds) and concluded that the CACC technology has the potential to significantly increase the highway capacity – potentially doubling the capacity at a high market penetration. Another important finding of this study was that the capacity effect is very sensitive to the market penetration, based on the fact that the reduced time gaps are only achievable between pairs of vehicles that are equipped with the CACC technology. CACC allows vehicles to communicate together and share important traffic information by transmitting information about incidents, speeds, positions, destinations, congestions from preceding vehicles, and emergencies [6]. Cooperative Following (CF) uses automated longitudinal control combined with inter-vehicle wireless communication allowing equipped vehicles to anticipate sudden and severe braking [7]. CF reduces shockwaves and smoothens the traffic flow and also improves the traffic safety.

The following two conclusions can be drawn from the literature review:

1- As the high rate of ACC market penetration could affect the traffic flow negatively, the research must focus more on the CACC rather than the ACC technology.

2- Extensive research is needed to study the effect of CACC on the freeway traffic flow in terms of throughput, congestion reduction, and capacity.

3. PROBLEM STATEMENT

With the occurring traffic congestions, there is a need to better manage the use of existing infrastructure in order to make the highway system more efficient and sustainable for the twenty-first century. The approach of our study is to better manage the existing capacity rather than looking into different ways of how to increase it. From the literature and our extensive observation of traffic congestions, one of the main causes of traffic on freeways is the continuing shockwaves, resulting from sudden vehicles braking, which create oscillations resulting in small traffic jams. With a limited highway capacity compared to the number of vehicles in operation, the small traffic jams ultimately become bottlenecks resulting in bumper-to-bumper traffic.

In this paper, we present an agent-based model of traffic congestion and examine the impact of having multiple CACC embedded vehicle(s) on a highway system consisting of 4 traffic lanes without overtaking. In a previous study [8], we worked on a model with a single lane and a single CACC car examining its impact on the traffic flow. This study is a continuation of the latter study. We examine the impact of having multiple CACC cars on 4 separate lanes with different traffic patterns (between cars, trucks, and CACC vehicles). In our agent-based model, CACC vehicles adapt their acceleration/deceleration according to vehicle-to-vehicle inter-communication. We analyze the average speed of the cars, the shockwaves, the evolution of traffic congestion, and other parameters throughout the lifecycle of the model. Failures in the operation of the sensors and communication equipment were not taken into consideration.

4. SIMULATION MODEL

The object-oriented model is developed using Java® general-purpose language. The microscopic traffic simulation model built is an expansion of an open-source model originally developed by [9]. Some of the most important additions to the original model are: adding two additional lanes, increasing the freeway distance from 2.5 Km to 10 Km, collecting macroscopic properties from the model as well as microscopic properties, and most importantly adding CACC vehicles.

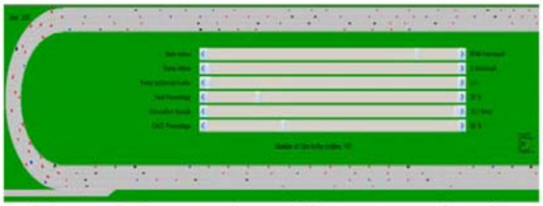


Figure 1. Screenshot of the proposed traffic simulator

As shown in fig. 1, our model consists of a U-shaped 4-laned freeway where ongoing traffic flows anticlockwise where vehicles enter and exit the system after traveling a distance of 10 km. Lane change, overtaking, and weaving are not allowed among vehicles, at this stage of our study. Cars (whether CACC or not) are 4 x 2 meters and trucks are 6 x 2 meters. The range of longitudinal detection of CACC sensors is 30 meters taken from [10]. At this stage, the model proposed does not attempt to find the optimal configuration (or calibration) of the CACC parameters. The proposed model uses the following CACC algorithm:

Phase 1 pseudo code

```
If [speed] of precedent_vehicle < [speed] of CACC_car AND separating_distance <= 14 meters AND [lane] of precedent_vehicle == [lane] of CACC_car
{
  Reduce [speed] of CACC_car by x%
}
Else
{
  Increase [speed] of CACC_car by x%
}
```

Phase 2 pseudo code

```
If [type] of precedent_vehicle == "CACC" AND [lane] of precedent_vehicle == [lane] of CACC_car
{
  Perform close following // meaning join platoon with a time gap of 0.5 s
}
```

The algorithm is divided into two phases. The first phase, which is the ACC part of the CACC system, is strictly related to the vehicle directly in front of the CACC vehicle. In this phase, the intelligent car communicates with the car in front of it and adapts its speed according to (a) the separating distance, (b) the lane used, and (c) the speed of the precedent vehicle. If the precedent vehicle was also CACC, close following (i.e. time gap of 0.5 s) can be achieved. Thus a platoon is formed in phase 2. If the precedent vehicle was non-intelligent, the CACC vehicle acts like ACC and maintains a safety distance time gap of 1.4 s using the algorithm stated in phase 1.

In our proposed model, the following parameters were used:

- Speed – speed of the operating vehicles ranging from 0 km/hr (complete stop) to 30.6 ms/frame equivalent to 130km/hr (maximum possible speed for cars) and 26 ms/frame equivalent to 110 km/hr (maximum possible speed for trucks).
- Acceleration – the acceleration of the operating vehicles ranging from 0 to 1 ms/frame.

• Deceleration – the deceleration of the operating vehicles ranging from 0 to -1 ms/frame.
• CACC Percentage – the percentage of CACC vehicles operating on the freeway.
• Time gap – the clearing safety distance between vehicles in time distance
• Truck Percentage – the percentage of trucks operating on the freeway.
• Main Inflow – the arrival rate of vehicles flowing into the freeway.
• Simulation time – each time tick in the model represents one minute.

The basic configuration used in this paper is a four-lane highway with the following parameters:

- The low arrival rate is set to 2000 vehicles/hour
- The high arrival rate is set to 6500 vehicles/hour
- The truck percentage is set to 10% capable of reaching 100%
- The CACC vehicles percentage is set to 0% capable of reaching 100%
- The simulation speed is set to 6.9 times (normal speed). It can be reduced or increased by the modeler.

5. MODEL ASSUMPTIONS AND LIMITATIONS

In this section, we present the model assumptions and limitations:

- Failures in the operation of the sensing and communication equipment of CACC are not considered.
- Only CACC cars are considered. All trucks operating are considered non-intelligent.
- Collisions, work zones, and weather conditions, although they highly affect the traffic flow, are not implemented.
- The CACC system is automatically turned off when a lane change is occurring or when there are no vehicles within the sensor's range of detection. The system is turned back on under normal conditions.
- The model proposed, at this stage, does not attempt to find the optimal configuration of the acceleration and deceleration parameters of the intelligent vehicles.
- The proposed model does not allow overtaking at this point. The current model is still a work in progress.

6. EXPERIMENTS AND RESULTS

We conducted two scenarios in this study. The first one consists of a low arrival rate of 2000 vehicles/hour, and the second consists of a high arrival rate of 6500 vehicles/hour. In each scenario, we introduced different levels of CACC penetration levels (i.e. 0%, 20%, 60%, 80%, and 100%). The cars enter the system with an initial speed set to the maximum – 130 km/hr and the trucks enter with their maximum speed of 110 km/hr. As the vehicles travel anticlockwise and exit the system, we collect statistics and analyze quantitatively the performance of the traffic flow. The results are shown in table 1 and 2. The following statistics are collected: (1) Mean speed, (2) flow of traffic, (3) standard deviation of speed, (4) density, and (5) average time spent in the system by vehicles. The same distance of 10 km is used in all the scenarios. All the simulations were ran 5 times for 30 minutes each. We chose a small number of replications because the results of the simulations were analogous.

Table 1. Scenario 1 results

2000 v/hr Arrival Rate	Mean Speed (km/hr)	Flow (v/hr)	Travel Time (min)	Density (v/km)	Std Dev
0% CACC	120.40	1630.00	5.79	20.08	1.90
20% CACC	119.92	1611.33	5.83	20.12	1.87
60% CACC	120.48	1595.33	5.79	20.11	2.05
80% CACC	120.80	1314.67	5.77	19.86	2.19
100% CACC	121.74	1570.67	5.70	19.81	2.17

Table 2. Scenario 2 results

6500 v/hr Arrival Rate	Mean Speed (km/hr)	Flow (v/hr)	Travel Time (min)	Density (v/km)	Std Dev
0% CACC	85.51	1329.33	7.09	74.70	7.95
20% CACC	84.69	1292.67	7.03	73.47	8.09
60% CACC	98.44	3828.67	6.60	75.01	5.17
80% CACC	106.49	4430.00	6.38	72.98	3.10
100% CACC	107.53	4745.33	6.31	72.31	3.78

In the low arrival rate scenario, the impact of CACC is minimal and leading to a modest loss of highway capacity and performance. Such negative results are not taken critically into account because of the minimal effect of CACC, and because of the fact that at a low arrival rate, the traffic performance is well in all cases.

Most of the study focuses on the high arrival rate scenario because this is the nature of the problem we're dealing with. As CACC vehicles enable closer vehicle following reaching a time gap of 0.5 seconds, significant highway capacity increases have been observed. Figure 2 shows the comparison of different CACC highway penetration in the case of traffic flow performance metric. Figure 3 and fig. 4, show the density and average speed at different CACC penetration levels. We can easily observe how the impact is directly related to the highway penetration. For instance, at 0% CACC penetration, the flow of traffic was slightly better than the case with 20% CACC penetration. With 60% CACC penetration, the average flow achieved a major progress reaching 3828.67 vehicles/hour. Evidently, the impact of CACC vehicles at higher penetration rate is positive. An improvement in all the performance measures collected was observed. At a rate of 0% CACC the flow averaged at 1329.33 vehicles/hour while it averaged at 4745.33 in 100% CACC penetration (almost 3.5 times more). Note here that by 100% CACC penetration, we mean 100% penetration as far as cars. The percentage of trucks (10%) did not change in all the scenarios. With a 100% CACC penetration including trucks the CACC impact can be even better reaching an average flow of 4860 vehicles/hour.

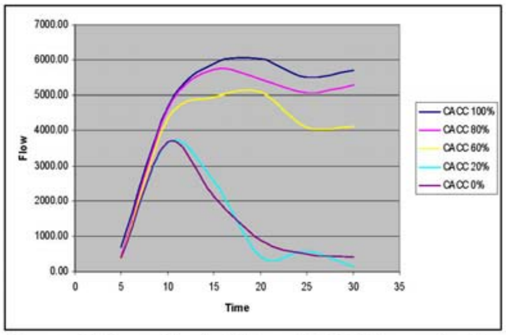


Figure 3. Density at different CACC penetration levels

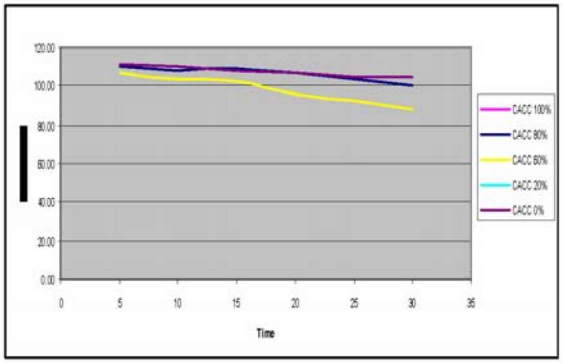


Figure 4. Average speed at different CACC penetration levels

7. CONCLUSION

In this paper, we identified the benefits of CACC and its impact on the traffic flow, especially in traffic congestions. We contributed a working prototype of a microscopic traffic simulator that uses macroscopic and microscopic traffic properties to describe the traffic performance. Four main conclusions were drawn from this study:

- At low traffic arrival rates, the impact of CACC is not significant. In some cases, the CACC impact resulted in a modest decrease in traffic performance. Therefore, it is suggested to turn off the CACC system at low traffic arrival rate situations, and activate the ACC system, to provide the driver with comfort and safety without affecting the traffic flow.

- At high traffic arrival rates, the impact of CACC is significantly positive in the cases of a highway penetration higher than 60%. An improvement in all the performance measures collected was observed.

-The CACC impact on traffic flow is directly related to the market penetration. A higher rate of penetration have a significant potential of increasing the highway capacity, while a low rate penetration could have no significant impact or even result in a modest loss of highway capacity.

-Since the CACC high penetration is not occurring in the near future, there is a need for a progressive deployment strategy, in order to increase the benefit of CACC. One approach, previously suggested by [6], suggests placing CACC vehicles in the same lane by giving them priority access to special lanes (such as HOV lanes). However, this approach has not been validated by any simulation studies.

- This study validated conclusions from two previous studies [1] and [6], that concluded that the impact of CACC at a high market penetration is significantly positive.

8. FUTURE WORK

Simulating traffic flow without overtaking is not realistic, since lane changes have a major impact on the traffic performance. Therefore, in our planned further work, overtaking will be implemented (i.e. cars will perform lane changing and weaving). In addition, the model parameters will be calibrated according to real life empirical freeway data. The CACC acceleration/deceleration parameter (i.e. the x% found in the pseudo code) will be optimized. Finally, as the CACC agents perform platoons, the platoons must be controlled for merging safety. Therefore, we will perform "controlled platooning" meaning that the number of platooning will be limited for more safety in merges (refer to [1])

REFERENCES

1. Bart van Arem, *M., IEEE*, Cornelie J. G. van Driel, and Ruben Visser, *The Impact of Cooperative Adaptive Cruise Control on Traffic-Flow Characteristics.* IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, 2006. **7**(4).
2. Payne, H.J., *FREFLO: A MACROSCOPIC SIMULATION MODEL OF FREEWAY TRAFFIC.* Transportation Research Board (TRB), 1979: p. p. 68-77.
3. Arem, P.J.Z.a.B.v., *Traffic effects of automated vehicle guidance system. A literature survey*, in *TNO Inro*, R. INRO-VVG, Editor. 1997: Delft, The Netherlands.
4. Hoetink, A.E., *Advanced Cruise Control en verkeersveiligheid*, in *R-2003-24*. 2003, Inst. for Road Safety Res. (SWOV): Leidschendam, The Netherlands.
5. B. van Arem, J.H.H., M. J. W. A. Vanderschuren, and C. H. Verheul, *An assessment of the impact of autonomous intelligent cruise control*, in *INRO-VVG, Report*, T. Inro, Editor. 1995: Delft, The Netherlands.
6. J. VanderWerf, S.E.S., M. A. Miller, and N. Kourjanskaia, *Evaluation of the effects of adaptive cruise control systems on highway traffic flow capacity and implications for deployment of future automated systems.* 81st Annual Meeting of the Transportation Research Board 2003: p. 78–84.
7. B. van Arem, C.M.J.T., and K. M. Malone,. *Modeling traffic flows with intelligent cars and intelligent roads*. in *Proc. IEEE Intell. Vehicles Symp*. 2003. Columbus, OH.
8. Georges M. Arnaout, M.T.K., Jun Zhang, Shannon R. Bowling. *An IntelliDrive Application for Reducing Traffic Congestions using Agent-Based approach*. in *Proceedings of the 2010 IEEE Systems and Information Engineering Design Symposium*. 2010. University of Virginia - Charlottesville, VA.
9. Treiber, M. *Microsimulation of Road Traffic*. 2010; Available from: http://www.traffic-simulation.de/.
10. Ann Hsu, F.E., Sonia Sachs, Pravin Varaiya, *The Design of Platoon Maneuver Protocols for IVHS*, in rt. *UCB-ITS-PRR-91-6*. 1991, PATH Research Repo

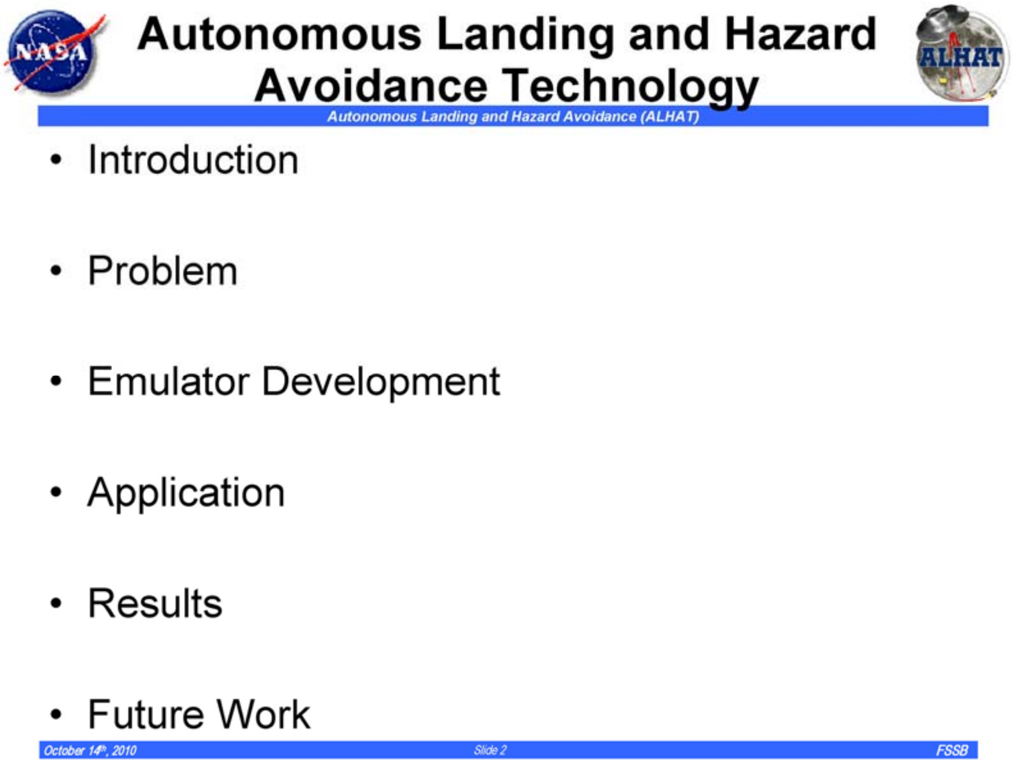
2.13 Flash LIDAR Emulator for HIL Simulation

Autonomous Landing and Hazard Avoidance (ALHAT)

Flash LIDAR Emulator for HIL Simulation

Paul Brewster
NASA Langley Research Center
ModSim World Conference
October 14th, 2010

October 14th, 2010 Slide 1 FSSB

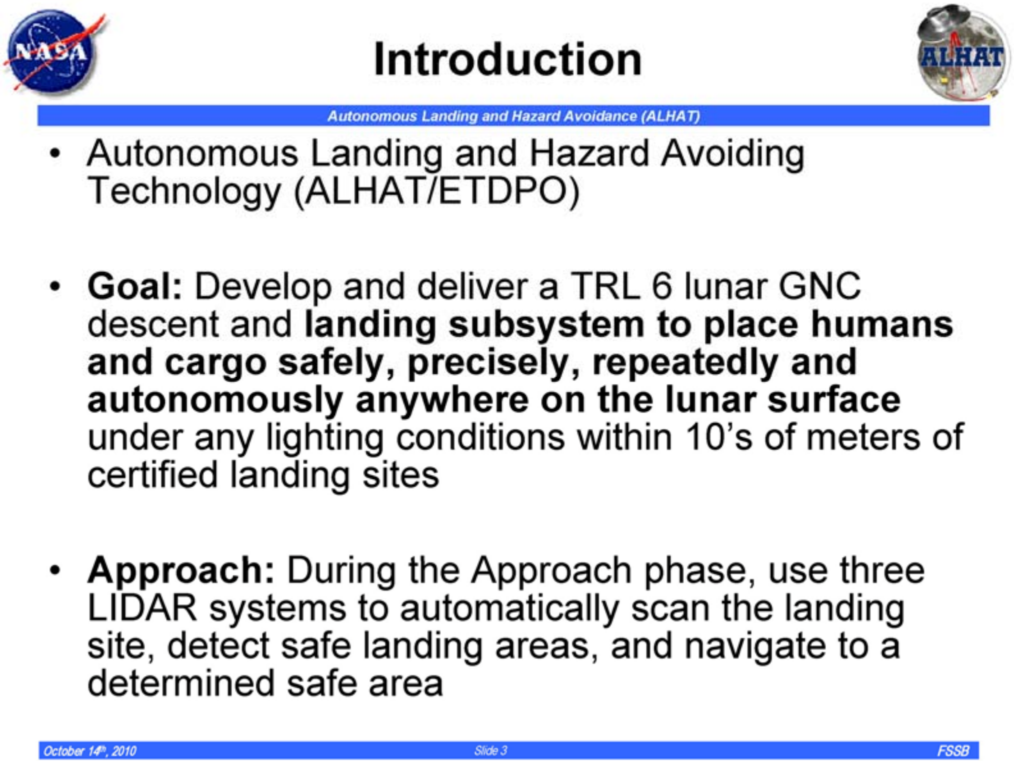


Autonomous Landing and Hazard Avoidance Technology

Autonomous Landing and Hazard Avoidance (ALHAT)

- Introduction
- Problem
- Emulator Development
- Application
- Results
- Future Work

October 14th, 2010 Slide 2 FSSB



Introduction

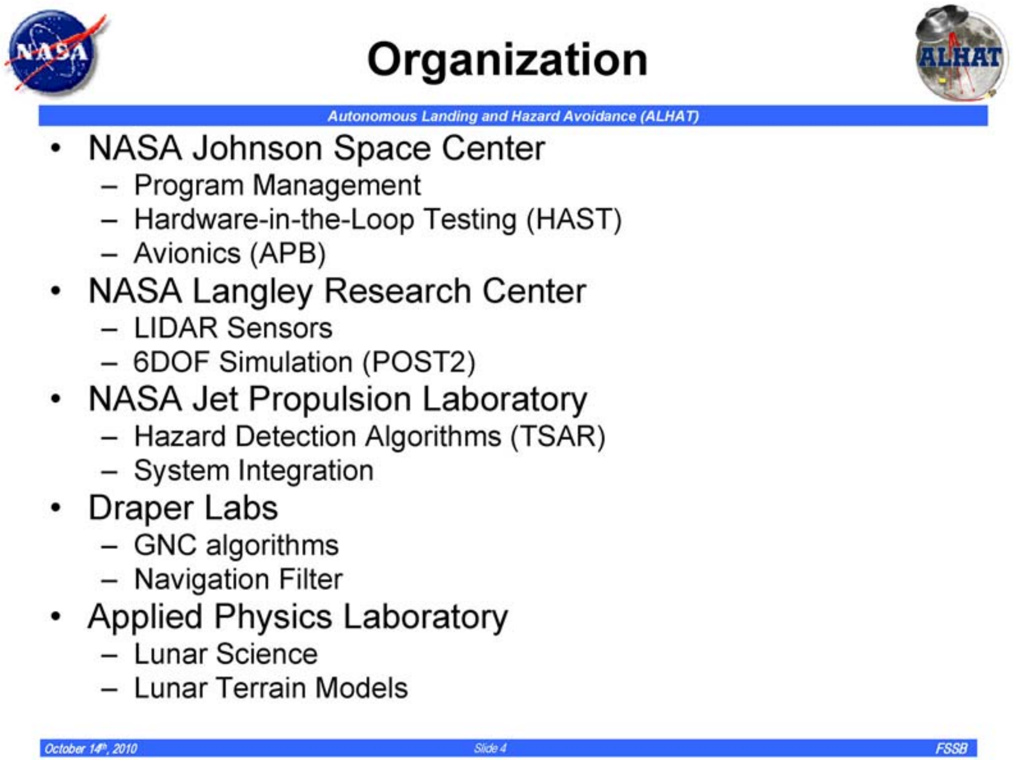
Autonomous Landing and Hazard Avoidance (ALHAT)

- Autonomous Landing and Hazard Avoiding Technology (ALHAT/ETDPO)

- **Goal:** Develop and deliver a TRL 6 lunar GNC descent and **landing subsystem to place humans and cargo safely, precisely, repeatedly and autonomously anywhere on the lunar surface** under any lighting conditions within 10's of meters of certified landing sites

- **Approach:** During the Approach phase, use three LIDAR systems to automatically scan the landing site, detect safe landing areas, and navigate to a determined safe area

October 14th, 2010 *Slide 3* *FSSB*

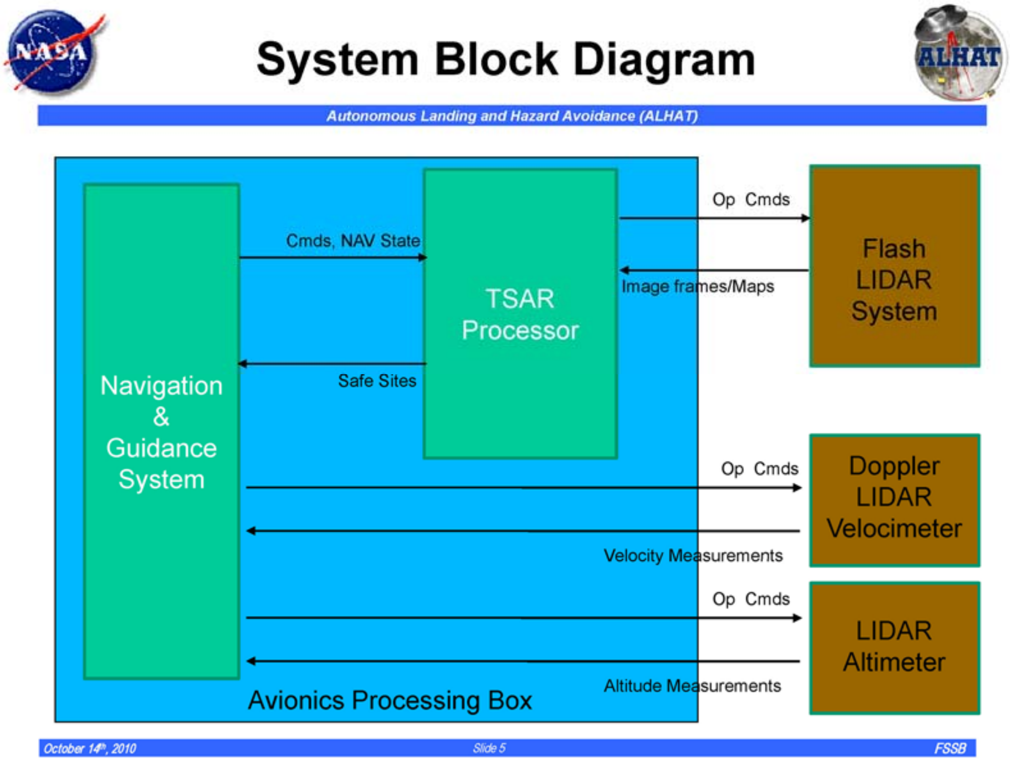


Organization

Autonomous Landing and Hazard Avoidance (ALHAT)

- NASA Johnson Space Center
  - Program Management
  - Hardware-in-the-Loop Testing (HAST)
  - Avionics (APB)
- NASA Langley Research Center
  - LIDAR Sensors
  - 6DOF Simulation (POST2)
- NASA Jet Propulsion Laboratory
  - Hazard Detection Algorithms (TSAR)
  - System Integration
- Draper Labs
  - GNC algorithms
  - Navigation Filter
- Applied Physics Laboratory
  - Lunar Science
  - Lunar Terrain Models

October 14th, 2010 *Slide 4* *FSSB*



System Block Diagram

Autonomous Landing and Hazard Avoidance (ALHAT)

Avionics Processing Box

Navigation & Guidance System

TSAR Processor

Flash LIDAR System

Doppler LIDAR Velocimeter

LIDAR Altimeter

Cmds, NAV State

Safe Sites

Op Cmds

Image frames/Maps

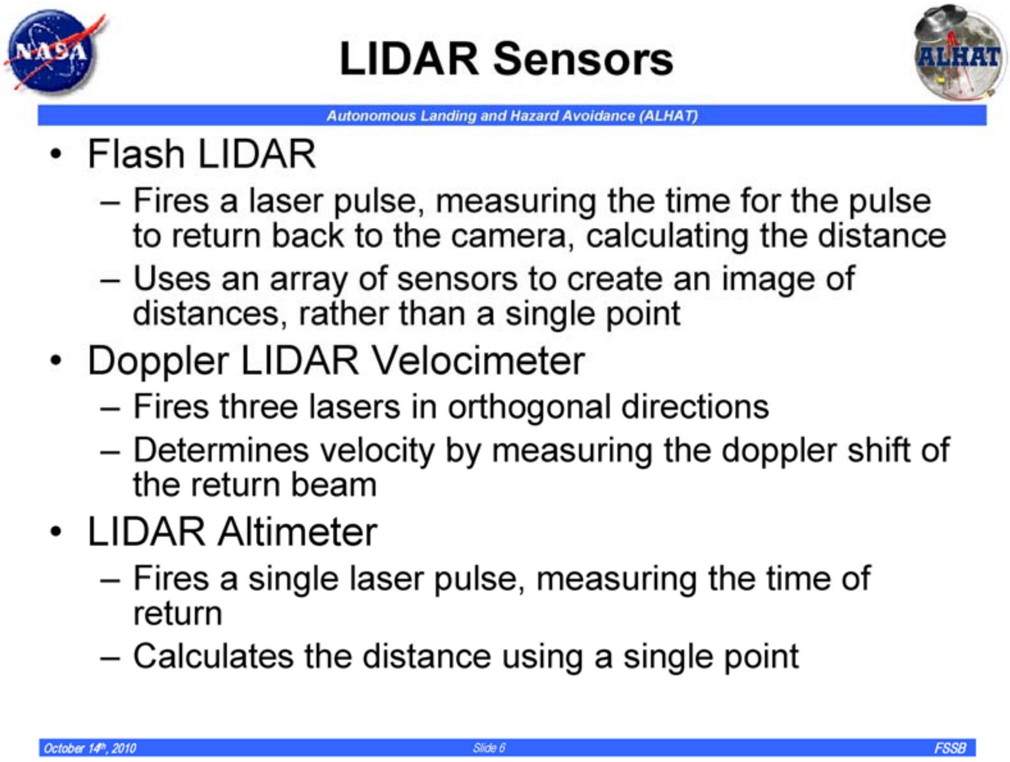
Op Cmds

Velocity Measurements

Op Cmds

Altitude Measurements

October 14th, 2010 Slide 5 FSSB

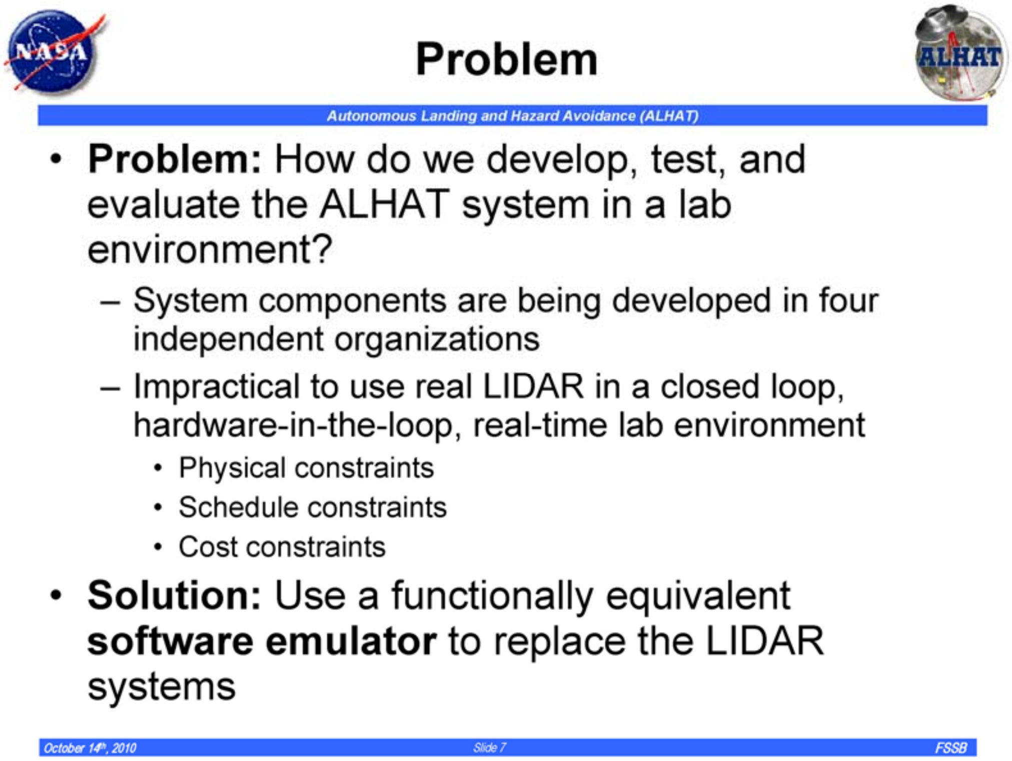


LIDAR Sensors

Autonomous Landing and Hazard Avoidance (ALHAT)

- Flash LIDAR
  - Fires a laser pulse, measuring the time for the pulse to return back to the camera, calculating the distance
  - Uses an array of sensors to create an image of distances, rather than a single point
- Doppler LIDAR Velocimeter
  - Fires three lasers in orthogonal directions
  - Determines velocity by measuring the doppler shift of the return beam
- LIDAR Altimeter
  - Fires a single laser pulse, measuring the time of return
  - Calculates the distance using a single point

October 14th, 2010 Slide 6 FSSB

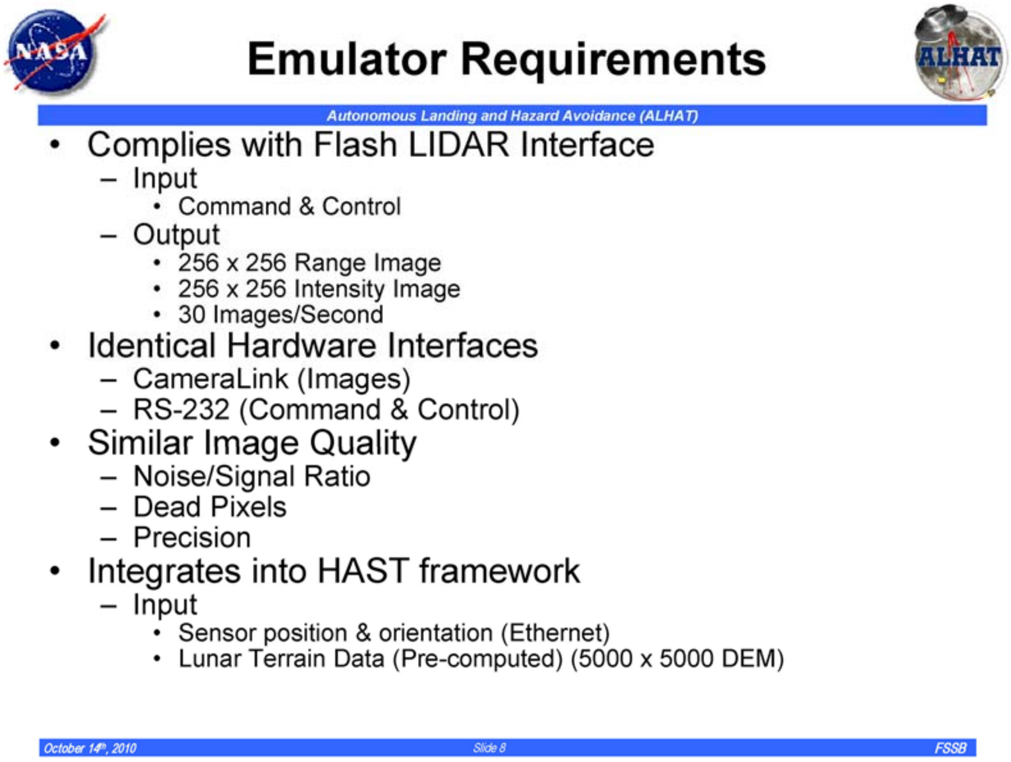


Problem

Autonomous Landing and Hazard Avoidance (ALHAT)

- **Problem:** How do we develop, test, and evaluate the ALHAT system in a lab environment?
  - System components are being developed in four independent organizations
  - Impractical to use real LIDAR in a closed loop, hardware-in-the-loop, real-time lab environment
    - Physical constraints
    - Schedule constraints
    - Cost constraints
- **Solution:** Use a functionally equivalent **software emulator** to replace the LIDAR systems

October 14th, 2010 Slide 7 FSSB



Emulator Requirements

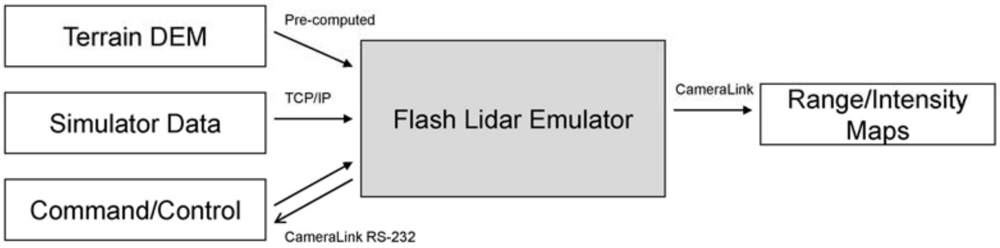
Autonomous Landing and Hazard Avoidance (ALHAT)

- Complies with Flash LIDAR Interface
  - Input
    - Command & Control
  - Output
    - 256 x 256 Range Image
    - 256 x 256 Intensity Image
    - 30 Images/Second
- Identical Hardware Interfaces
  - CameraLink (Images)
  - RS-232 (Command & Control)
- Similar Image Quality
  - Noise/Signal Ratio
  - Dead Pixels
  - Precision
- Integrates into HAST framework
  - Input
    - Sensor position & orientation (Ethernet)
    - Lunar Terrain Data (Pre-computed) (5000 x 5000 DEM)

October 14th, 2010 Slide 8 FSSB

Emulator Interfaces

Autonomous Landing and Hazard Avoidance (ALHAT)



Terrain DEM

Simulator Data

Command/Control

Pre-computed

TCP/IP

CameraLink RS-232

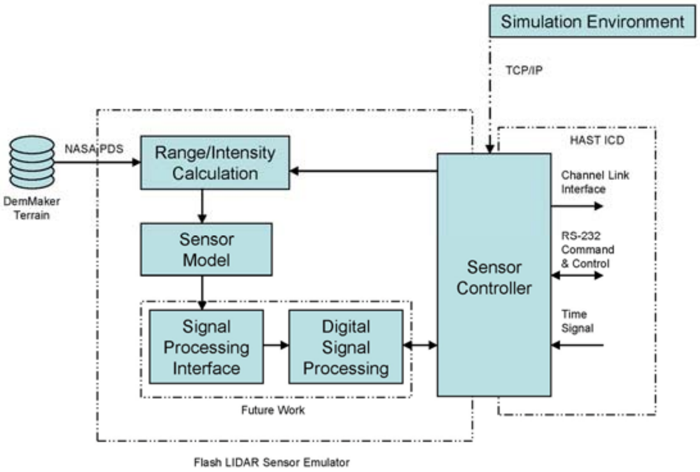
Flash Lidar Emulator

CameraLink

Range/Intensity Maps

Emulator Block Diagram

Autonomous Landing and Hazard Avoidance (ALHAT)



Simulation Environment

TCP/IP

HAST ICD

DemMaker Terrain

NASA PDS

Range/Intensity Calculation

Sensor Model

Signal Processing Interface

Digital Signal Processing

Future Work

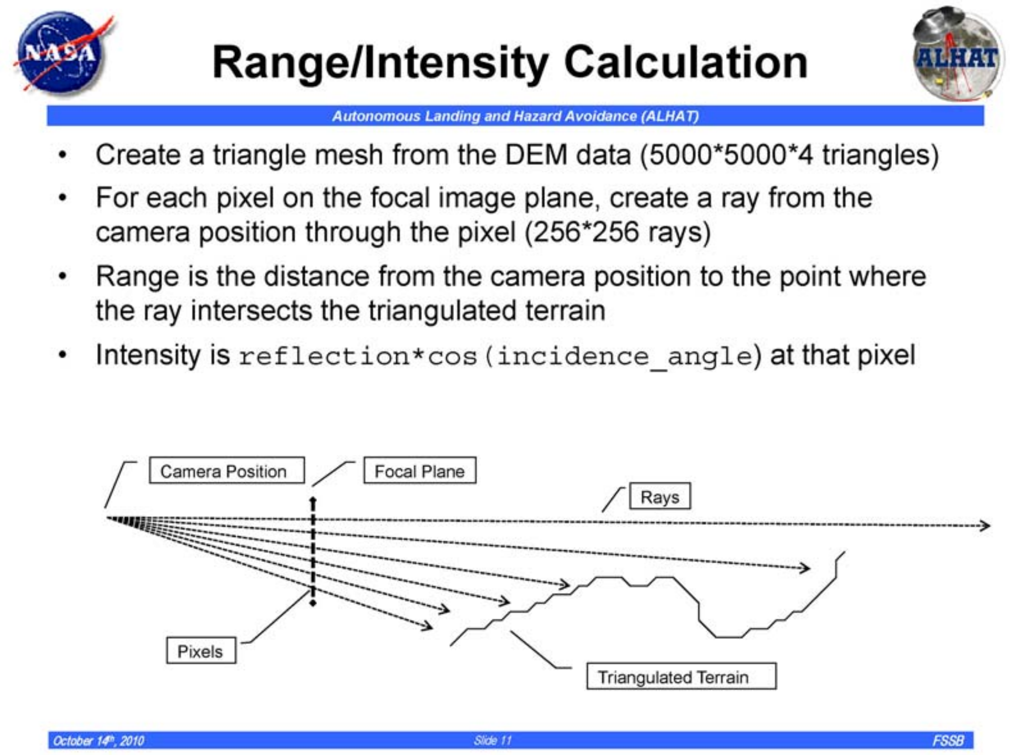
Sensor Controller

Channel Link Interface

RS-232 Command & Control

Time Signal

Flash LIDAR Sensor Emulator

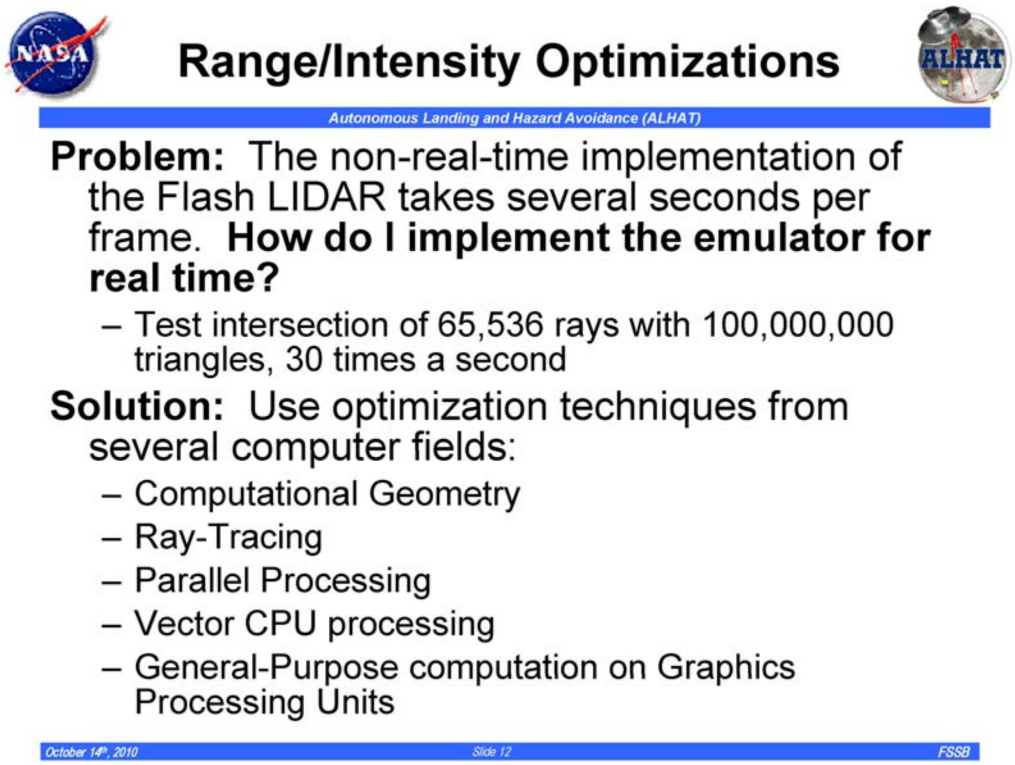


Range/Intensity Calculation

Autonomous Landing and Hazard Avoidance (ALHAT)

- Create a triangle mesh from the DEM data (5000*5000*4 triangles)
- For each pixel on the focal image plane, create a ray from the camera position through the pixel (256*256 rays)
- Range is the distance from the camera position to the point where the ray intersects the triangulated terrain
- Intensity is `reflection*cos(incidence_angle)` at that pixel

October 14th, 2010 Slide 11 FSSB



Range/Intensity Optimizations

Autonomous Landing and Hazard Avoidance (ALHAT)

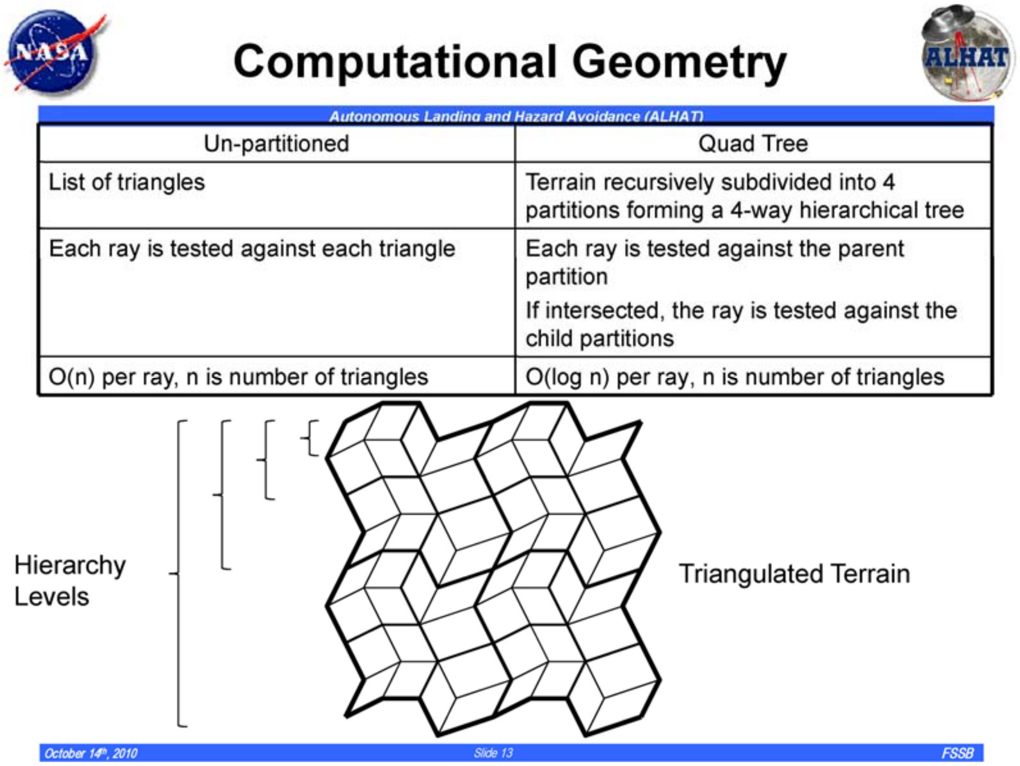
Problem: The non-real-time implementation of the Flash LIDAR takes several seconds per frame. **How do I implement the emulator for real time?**

- Test intersection of 65,536 rays with 100,000,000 triangles, 30 times a second

Solution: Use optimization techniques from several computer fields:

- Computational Geometry
- Ray-Tracing
- Parallel Processing
- Vector CPU processing
- General-Purpose computation on Graphics Processing Units

October 14th, 2010 Slide 12 FSSB

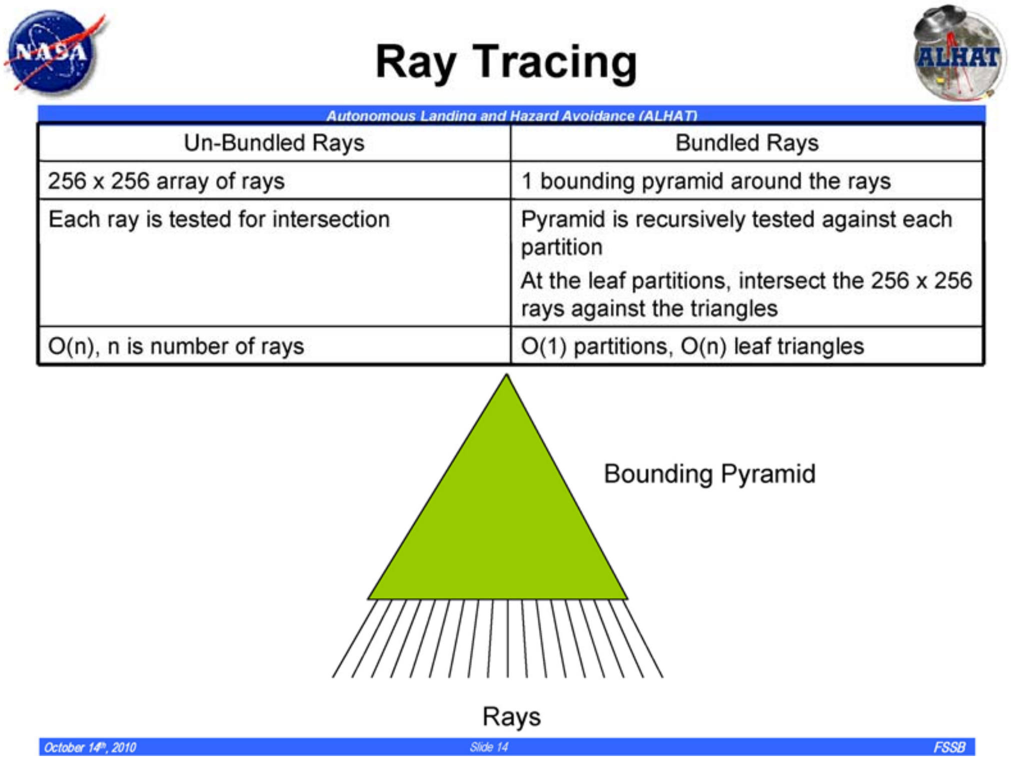


Computational Geometry

Autonomous Landing and Hazard Avoidance (ALHAT)

Un-partitioned	Quad Tree
List of triangles	Terrain recursively subdivided into 4 partitions forming a 4-way hierarchical tree
Each ray is tested against each triangle	Each ray is tested against the parent partition If intersected, the ray is tested against the child partitions
O(n) per ray, n is number of triangles	O(log n) per ray, n is number of triangles

October 14th, 2010 Slide 13 FSSB



Ray Tracing

Autonomous Landing and Hazard Avoidance (ALHAT)

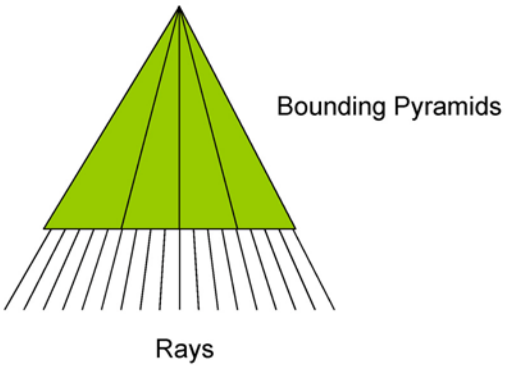
Un-Bundled Rays	Bundled Rays
256 x 256 array of rays	1 bounding pyramid around the rays
Each ray is tested for intersection	Pyramid is recursively tested against each partition At the leaf partitions, intersect the 256 x 256 rays against the triangles
O(n), n is number of rays	O(1) partitions, O(n) leaf triangles

October 14th, 2010 Slide 14 FSSB

Parallel Processing

Autonomous Landing and Hazard Avoidance (ALHAT)

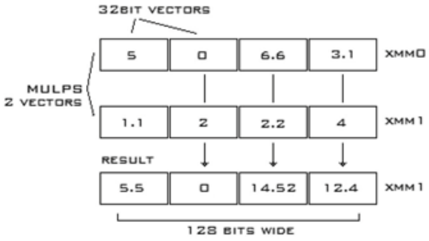
Single Bundle	Parallel Bundles
All the rays in a single bundle	Divide the bundle into sub-bundles, one for each CPU core
Not easy to parallelize	Independent tasks for 100% parallelization



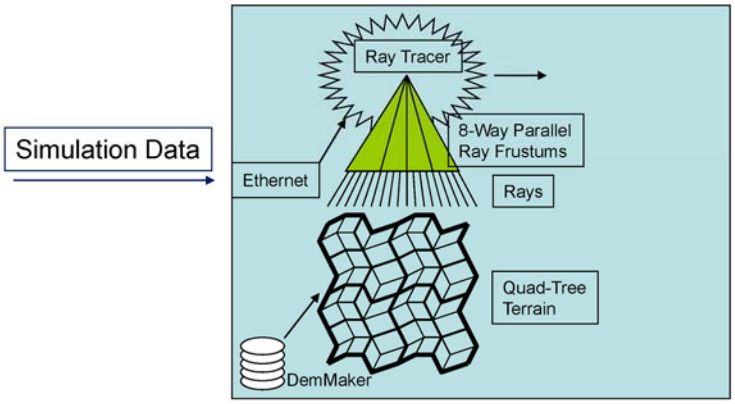
Vector CPU Processing

Autonomous Landing and Hazard Avoidance (ALHAT)

- Modern CPUs are scalar processors
  - Each instruction operates on one data item at a time
- Streaming SIMD Extensions
  - Intel extend the x86 instruction set (SSE)
  - One instruction can operate on
    - 4 32-bit integers
    - 4 32-bit floats
    - 2 64-bit floats
    - 2 64-bit integers
    - 8 16-bit integers
    - 16 8-bit characters
  - Ideal for Vector/Matrix math
  - Additional instructions that must be explicitly used



Emulator Design

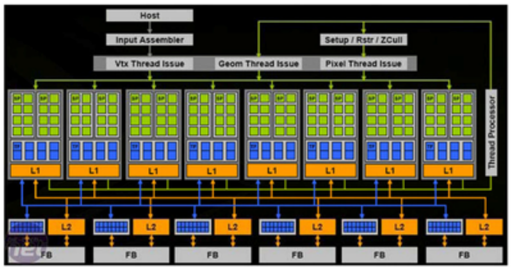


Sensor Model

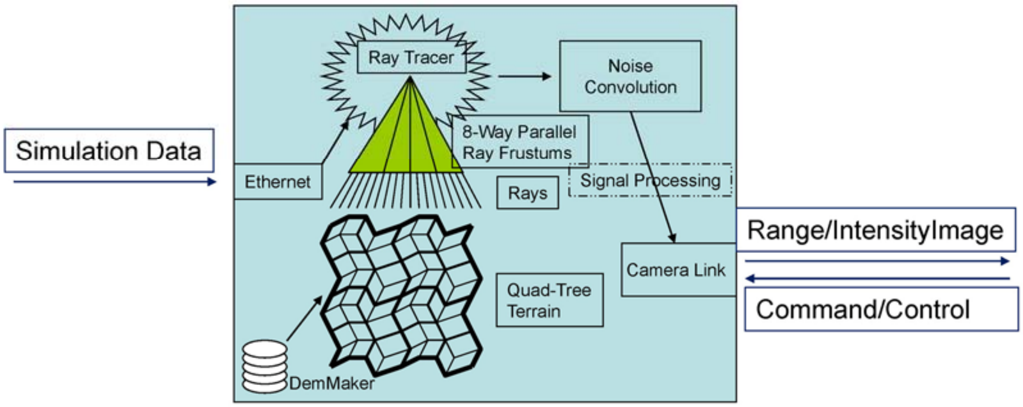
- Add Gaussian noise to each pixel in the image
  - Signal/Noise Ratio
  - Based on POST2 sensor model or actual hardware characteristics
- Pre-calculate random dead pixels
- Use intensity value for pixel cut-out
- Convolve with Gaussian filter for crosstalk or bleeding between pixels

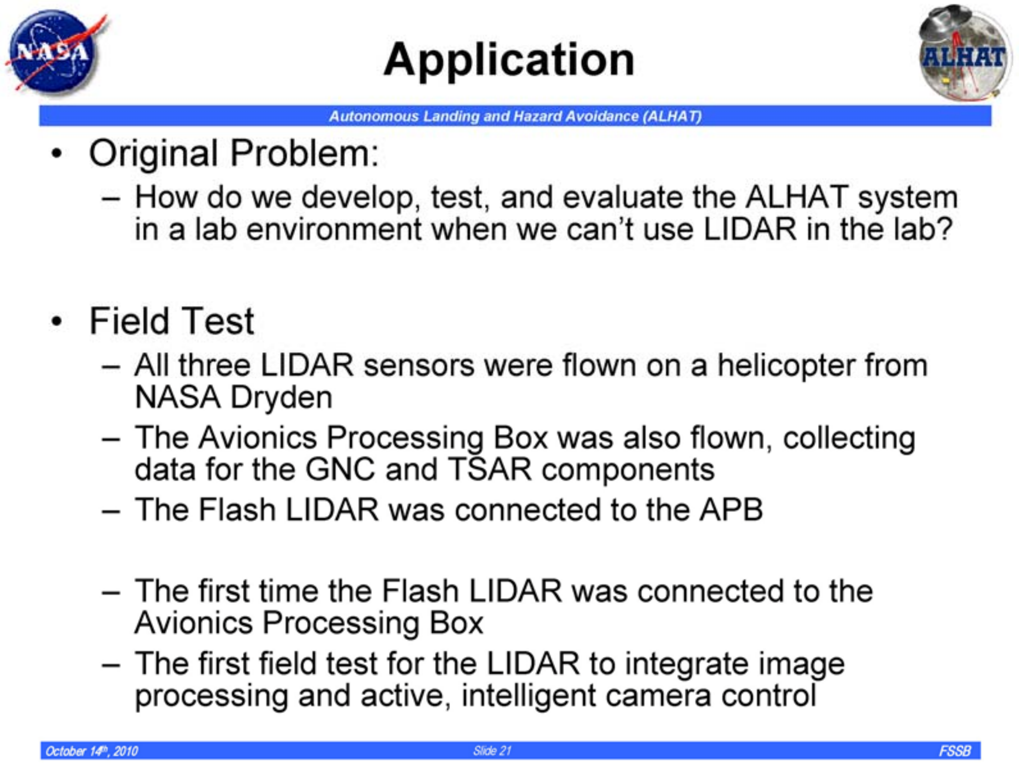
General-Purpose computation on Graphics Processing Units

- A modern GPU is bigger and has more computational power than the CPU
- Massively parallel, multi-core processor
  - Hundreds of cores per processor
  - Each core is a vector processor
- Ideal for image processing
  - Each pixel will execute the exact same program, in parallel
- Implemented
  - Additive Gaussian Noise
  - Gaussian Convolution
  - Pixel Cut-Out
  - Image Formatting



Emulator Design



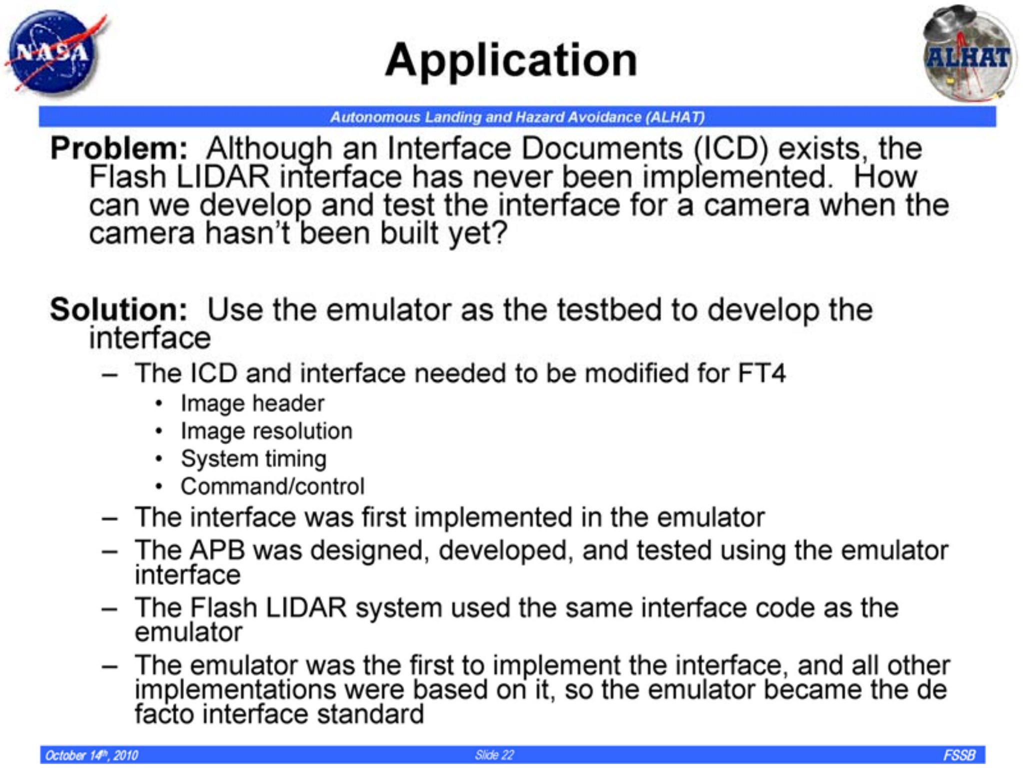


Application

- Original Problem:
  - How do we develop, test, and evaluate the ALHAT system in a lab environment when we can't use LIDAR in the lab?

- Field Test
  - All three LIDAR sensors were flown on a helicopter from NASA Dryden
  - The Avionics Processing Box was also flown, collecting data for the GNC and TSAR components
  - The Flash LIDAR was connected to the APB

  - The first time the Flash LIDAR was connected to the Avionics Processing Box
  - The first field test for the LIDAR to integrate image processing and active, intelligent camera control

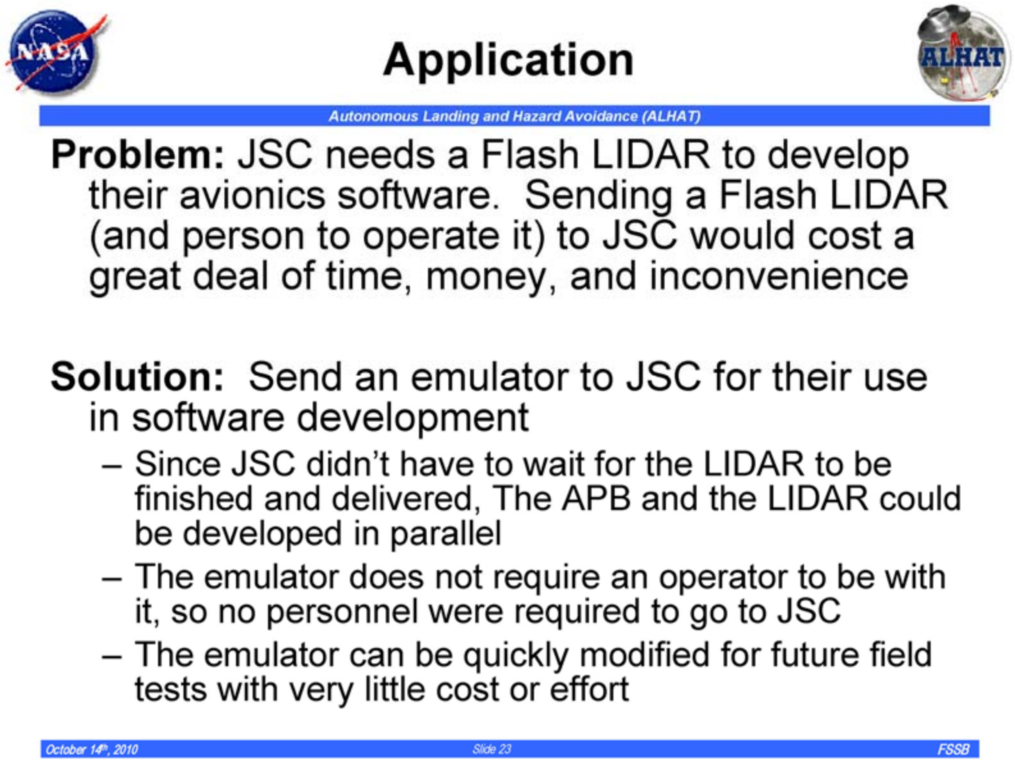


Application

Problem: Although an Interface Documents (ICD) exists, the Flash LIDAR interface has never been implemented. How can we develop and test the interface for a camera when the camera hasn't been built yet?

Solution: Use the emulator as the testbed to develop the interface

- The ICD and interface needed to be modified for FT4
  - Image header
  - Image resolution
  - System timing
  - Command/control
- The interface was first implemented in the emulator
- The APB was designed, developed, and tested using the emulator interface
- The Flash LIDAR system used the same interface code as the emulator
- The emulator was the first to implement the interface, and all other implementations were based on it, so the emulator became the de facto interface standard



Application

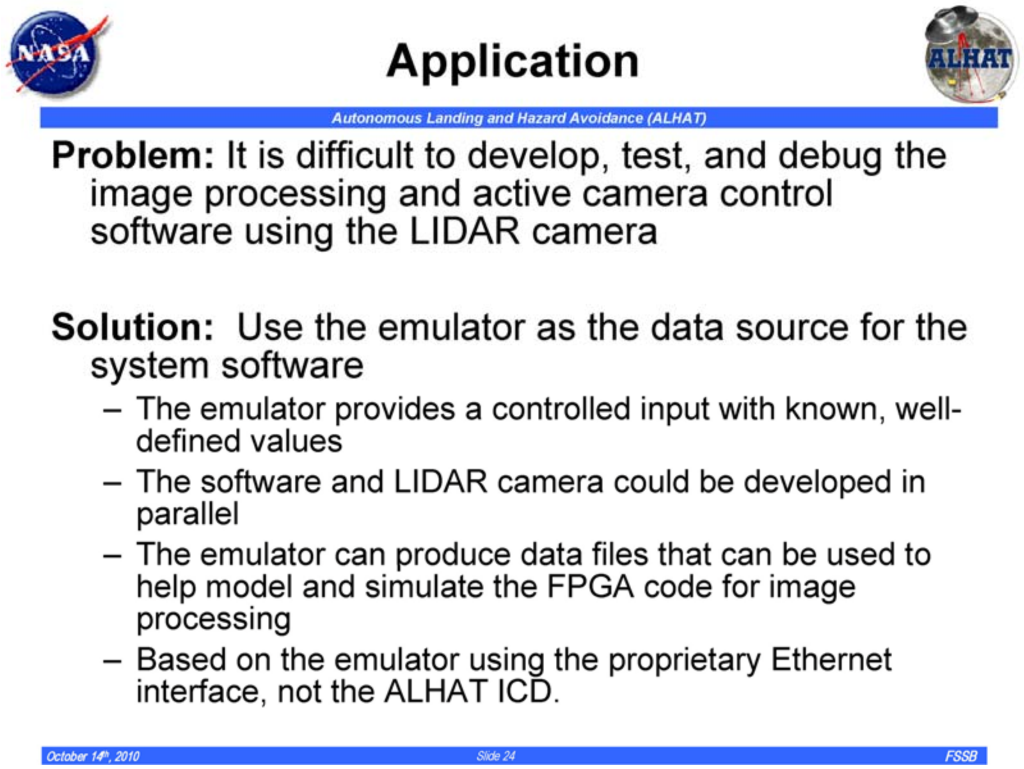
Autonomous Landing and Hazard Avoidance (ALHAT)

Problem: JSC needs a Flash LIDAR to develop their avionics software. Sending a Flash LIDAR (and person to operate it) to JSC would cost a great deal of time, money, and inconvenience

Solution: Send an emulator to JSC for their use in software development

- Since JSC didn't have to wait for the LIDAR to be finished and delivered, The APB and the LIDAR could be developed in parallel
- The emulator does not require an operator to be with it, so no personnel were required to go to JSC
- The emulator can be quickly modified for future field tests with very little cost or effort

October 14th, 2010 Slide 23 FSSB



Application

Autonomous Landing and Hazard Avoidance (ALHAT)

Problem: It is difficult to develop, test, and debug the image processing and active camera control software using the LIDAR camera

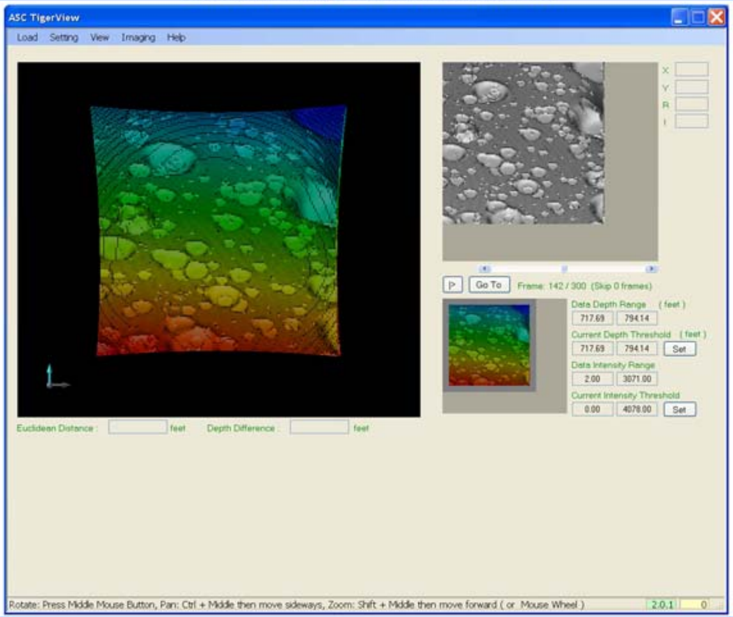
Solution: Use the emulator as the data source for the system software

- The emulator provides a controlled input with known, well-defined values
- The software and LIDAR camera could be developed in parallel
- The emulator can produce data files that can be used to help model and simulate the FPGA code for image processing
- Based on the emulator using the proprietary Ethernet interface, not the ALHAT ICD.

October 14th, 2010 Slide 24 FSSB

Results

Autonomous Landing and Hazard Avoidance (ALHAT)



Frame: 142 / 300 (Skip 0 frames)

Data Depth Range (feet): 717.69 794.14

Current Depth Threshold (feet): 717.69 794.14

Data Intensity Range: 2.00 3071.00

Current Intensity Threshold: 0.00 4078.00

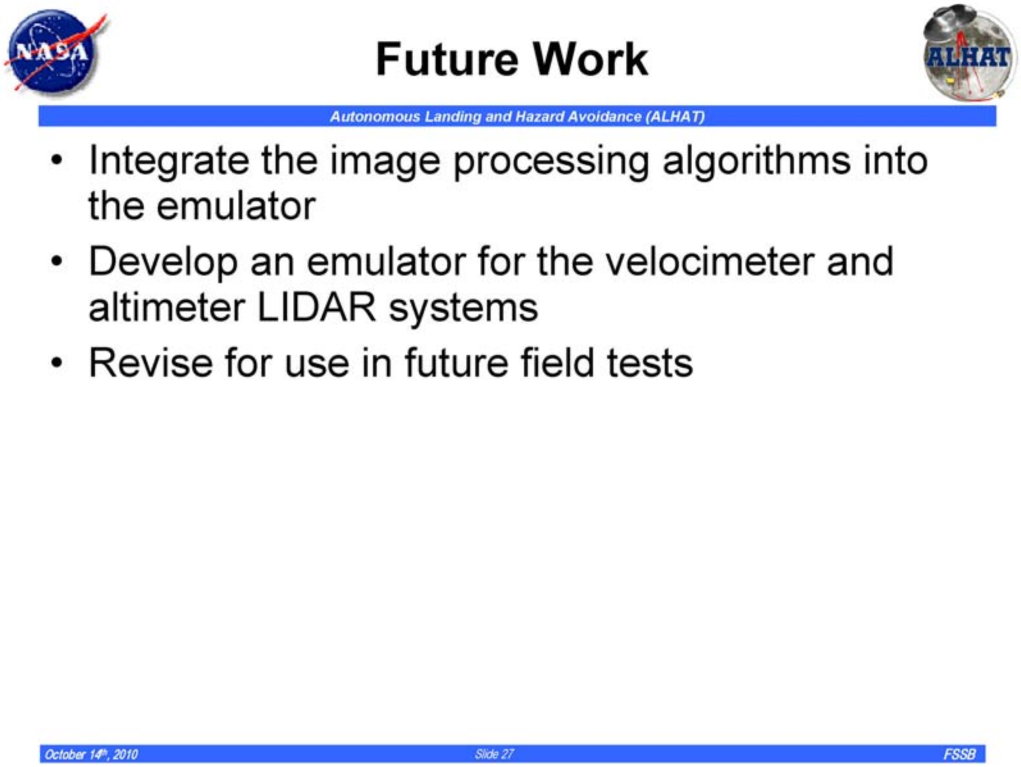
Euclidean Distance : feet Depth Difference : feet

Rotate: Press Middle Mouse Button, Pan: Ctrl + Middle then move sideways, Zoom: Shift + Middle then move forward (or Mouse Wheel) 2.0.1 0

Results

Autonomous Landing and Hazard Avoidance (ALHAT)

- An emulator was delivered to JSC in August 2009
- JSC used the emulator to develop their APB interface software in preparation for the flight test
- There were no significant interface issues in the flight test, despite the APB and the LIDAR never being physically connected until the field test
- An emulator was used extensively in the development process at Langley. All initial FPGA code was developed and tested using the emulator first
- An emulator was sent to the test site and used for debugging and integration leading up to the test



Future Work

Autonomous Landing and Hazard Avoidance (ALHAT)

- Integrate the image processing algorithms into the emulator
- Develop an emulator for the velocimeter and altimeter LIDAR systems
- Revise for use in future field tests

October 14th, 2010 Slide 27 FSSB

2.14 ARC+® & ARC PC Welding Simulators: Teach Welders with Virtual Interactive 3D Technologies

ARC+® & ARC PC Welding Simulators: Teach Welders with Virtual Interactive 3D Technologies

Claude Choquet[1]

[1] 123 Certification Inc, 1751 Richardson Street, #2204, Montreal, Quebec, Canada, H3K1G6

E-mail: cchoquet@123certification.com

Abstract

123 Certification Inc., a Montreal based company, has developed an innovative hands-on welding simulator solution to help build the welding workforce in the most simple way. The solution lies in virtual reality technology, which has been fully tested since the early 90's. President and founder of 123 Certification Inc., Mr. Claude Choquet Ing. Msc. IWE, acts as a bridge between the welding and the programming world. Working in these fields for more than 20 years, he has filed 12 patents world-wide for a gesture control platform with leading edge hardware related to simulation. In the summer of 2006, Mr Choquet was proud to be invited to the annual IIW International Welding Congress in Quebec City to launch the ARC+ welding simulator. A 100% virtual reality system and web based training center was developed to simulate multi-process, multi-material, multi-position, and multi-pass welding. The simulator is intended to train welding students and apprentices in schools or industries. The welding simulator is composed of a real welding electrode holder (SMAW-GTAW) and gun (GMAW-FCAW), a head-mounted display (HMD), a 6 degrees of freedom tracking system for interaction between the user's hands and head, as well as external audio speakers. Both guns and HMD are interacting online and simultaneously. The welding simulation is based on the law of physics and empirical results from detailed analysis of a series of welding tests based on industrial applications tested over the last 20 years.[1]

The simulation runs in real-time, using a local logic network to determine the quality and shape of the created weld. These results are based on the orientation, distance, and speed of the welding torch and depth of penetration. The welding process and resulting weld bead are displayed in a virtual environment with screenplay interactive training modules. For review, weld quality and recorded process values can be displayed and diagnosed after welding. To help in the learning process, a learning curve for each student and each Virtual Welding Class® can be plotted, for an instructor's review or a required third party evaluation.

1. Introduction

By way of introduction, here is a quote from the senior editor of the US-based magazine The Fabricator:

"A recent study led to an odd conclusion: Playing video games may produce better surgeons."

Really, it's true, at least according to researchers at the Banner Health Center in Phoenix. They had surgical residents play games on the Nintendo® Wii™ console before simulated surgeries. Games such as Marble Mania require precise hand movements of the Wii's wireless wand, movements that seem to prep these residents for the hand movements surgery requires.

In a certain light, welding resembles surgery, with careful hands "stitching up" metal instead of skin..." [2]-[3].

2. Work accomplished

A welder's helmet and welding gun have been instrumented for the welding simulator ARC+. The electronic sensors and the head mounted device are installed in order to provide visibility during welding. It tracks the hand motions and process while the 3D glasses inside the Welder's Helmet provide high 3D graphics (Fig. 1).

A scientific breakthorugh enables now a welder to be trained remotely thanks to the new generations of PC mouses, For exemple since February 2010, ARC PC is now available for free demonstration and netmeetings fully remotely on any PC

A new approach to motion dexetrity memory training is now available at very low cost thanks to application of a list of Gameplay-Based Design Principles,

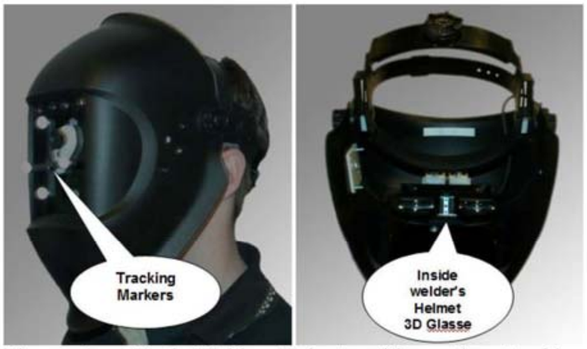


Figure 1: Real Welder Helmet with motion tracking technology Source: 123Certification Inc.

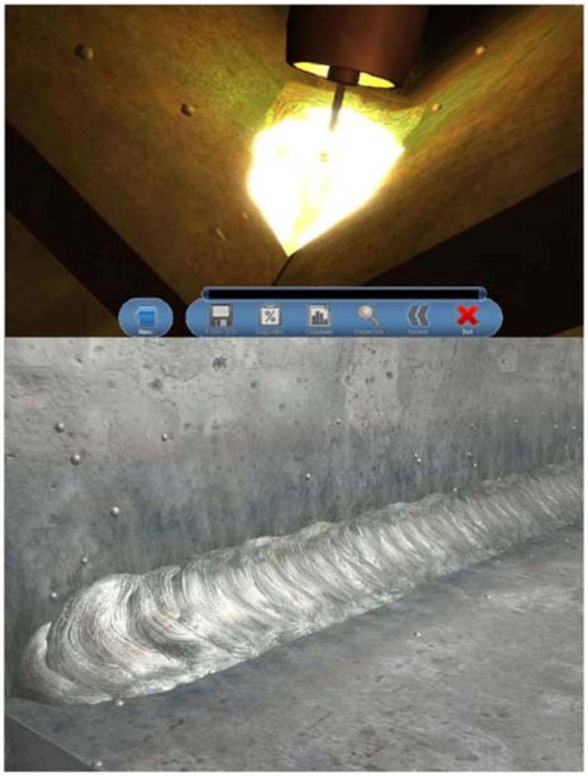


Figure 2: A virtual reality simulation of gas metal arc welding a fillet weld (GMAW). Source: 123 Certification

The solution is based on 100% virtual reality welding simulation without arc and metal work piece. High resolution images can be seen by the student on 3D glasses, and projected on a screen for classroom observation or replays. Under supervision, the welder can learn how to maintain a weld pool, arc length and consistent welding bead.. He can also learn how to cope with the 3 arc metal transfer mode (short-circuit, globular and arc spray), sparks and virtual fumes and volatiles. The database link to the simulator and the welder's motion activities reproduces the same results obtained in a welding booth.

As this technology evolves, increasing attention will be paid to the development of fine motor skills. For example, the simulator is able to track head and hand motions during a weld, helping the student determine the optimal angle of view while laying down a weld bead. Next research activities will be on image rendering of grain microstructure recrystallization after welding in the H.A.Z responsible for quality such as metallurgy based weld defects. (Fig. 2).

The ARC+ simulator is capable of all manual and semi-automatic welding processes (Fig. 3). At the time of this article, more than 50 welding data sheets and 1500 exercises had been implemented. An approximate 900 hours of exercises are available for perfecting fine motor skills. The possibilities of exercises are limitless. All welding data sheets with different processes, materials, welding positions, assembly preparations, multi-pass, weaving or linear are available.

This presentation will focus on light alloys requiring GTAW (TIG Welding) such as components submitted to heat from airplane reactors. (Fig. 4).

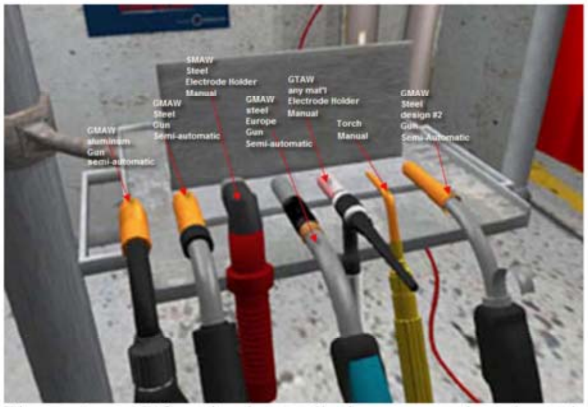
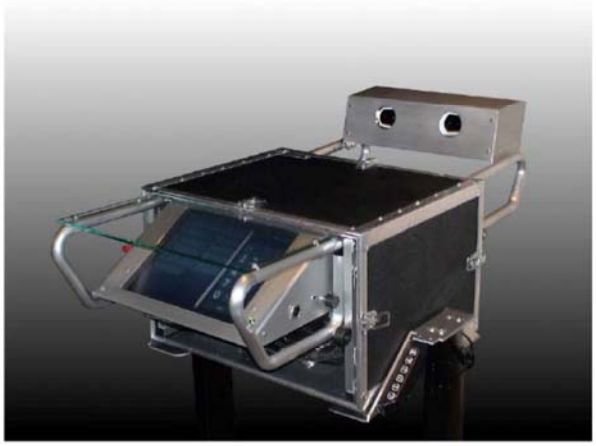


Figure 3: ARC+ simulates all these manual and semi-automatic welding processes Source: 123Certification Inc.



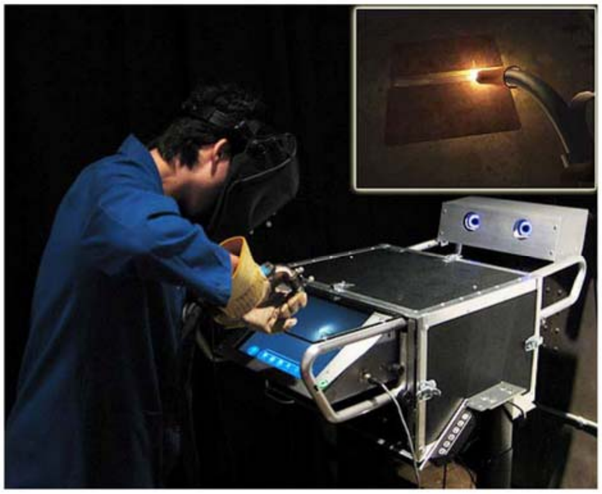


Figure 4: ARC+ welding simulator in action with the GTAW (TIG Welding) for a welder training.Source: 123Certification

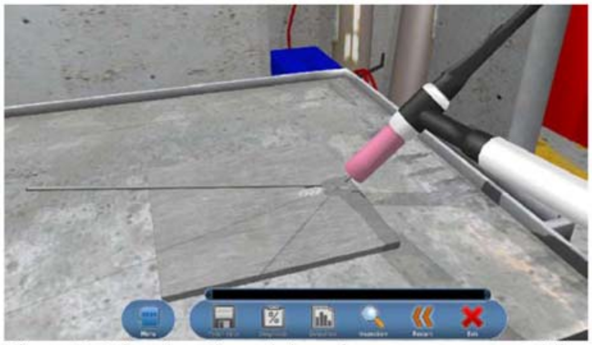


Figure 5: Virtual reality training of gas tungsten arc welding (GTAW). Source: 123 Certification Inc.

The ARC+ welding simulator allows the user to generate a virtual welder hands-on action environment with the help of patent pending technologies. It detects welder motion, processes equations between metallurgy and motion as well as casts a 3-D image of the user's gestures.

It recognizes some welding defects and their causes. While taking into account several variables affecting weld soundness, it gives the user an opportunity to evaluate his or her work with a diagnostic report, a ranking or a grade, and a visual examination as per a real welded assembly (Fig. 9). The diagnostic with a tolerance of 0.25mm RMS provides reports with motion dexterity accuracy results and some weld defects that allow the welder to take immediate corrective action, and to continuously improve his skills in a safe and enjoyable setting.

For example, we have broken down the hand motion into 5 essential variables: motion straightness, arc speed, stick-out, work angle and travel angle (Fig. 6).

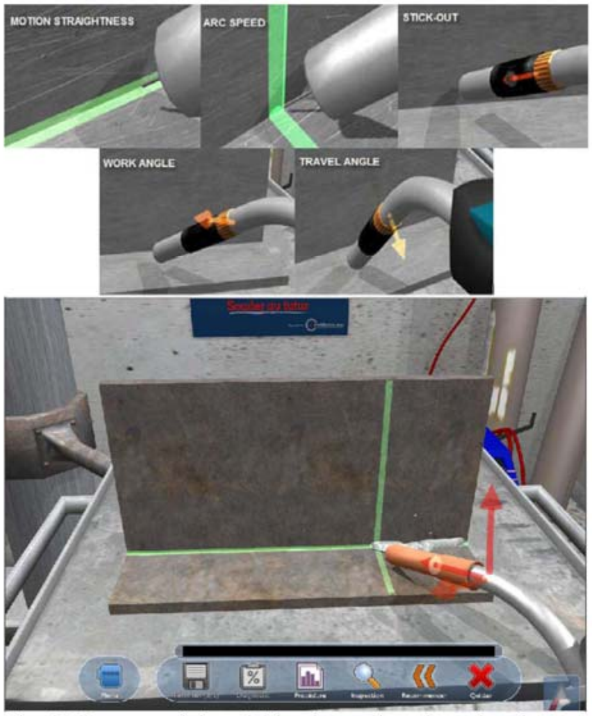
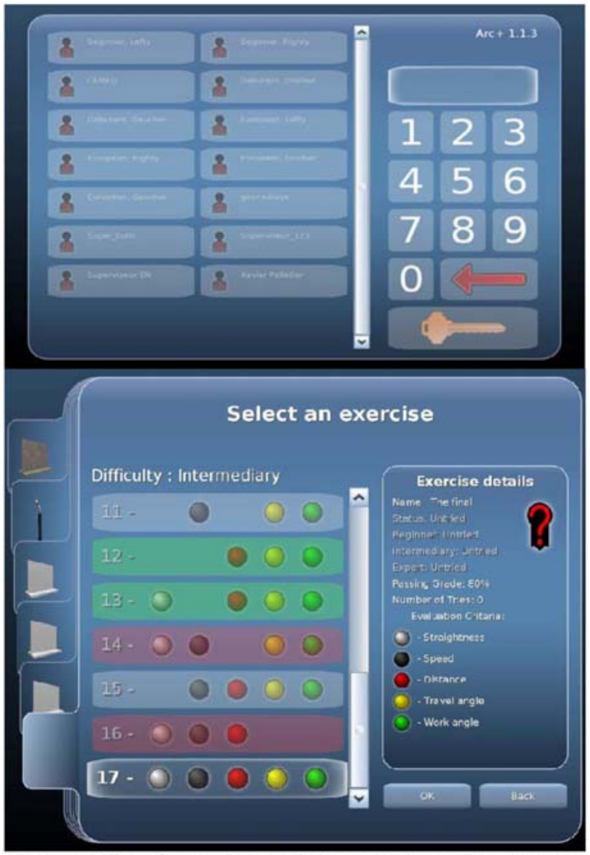


Fig 6 Five essential variables in motion dexterity that are tracked with the ARC+ simulator. Source: 123 Certification Inc.

The ARC+ simulator provides an atmosphere conducive to interactive learning thanks to numerous welding exercises based on the 5 motion dexterity variables. For example, 17 standard exercises (Fig. 7) based on those variables are available for all welding data experimentations.

In a guided training mode, a trainee would have to succeed in an exercise before stepping up to the next one. The ARC+ simulator also offers to the usual beginner, intermediate and expert, different level exercises requiring progressively more skills.



Select an exercise

Difficulty : Intermediary

11 -
12 -
13 -
14 -
15 -
16 -
17 -

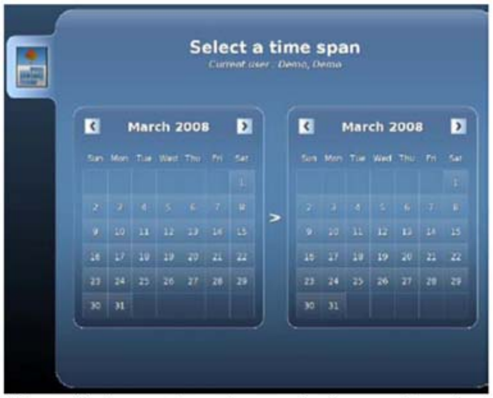
Exercise details

Name: The final
Status: Untried
Beginner: Untried
Intermediary: Untried
Expert: Untried
Passing Grade: 80%
Number of Tries: 0
Evaluation Criteria:

- ● - Straightness
- ● - Speed
- ● - Distance
- ● - Travel angle
- ● - Work angle

Figure 7 List of exercises based on the 5 motion dexterity essential variables. Source: 123Certification Inc.

In replay mode, an instructor can review the stick-out or the work angle in a precise location of a bead. He or she can help the student understand the weld size or the root penetration (Fig. 8). The reviews are not only for the diagnostic report but also for the motion dexterity tracked during welding. Replay can also have an unlimited angle of view to track a point on interest between the diagnostic and the welder's motion. The replay mode is a demonstration of previous trainee hands-on activities. The trainer can easily access all saved weld beads by the trainee. Trainer can easily identify and transfer to the trainee the hands-on results on the replay mode.



Select a time span

Current user : Demo, Demo

Figure 8 Access to replay mode for result reviews. Source: 123Certification Inc.

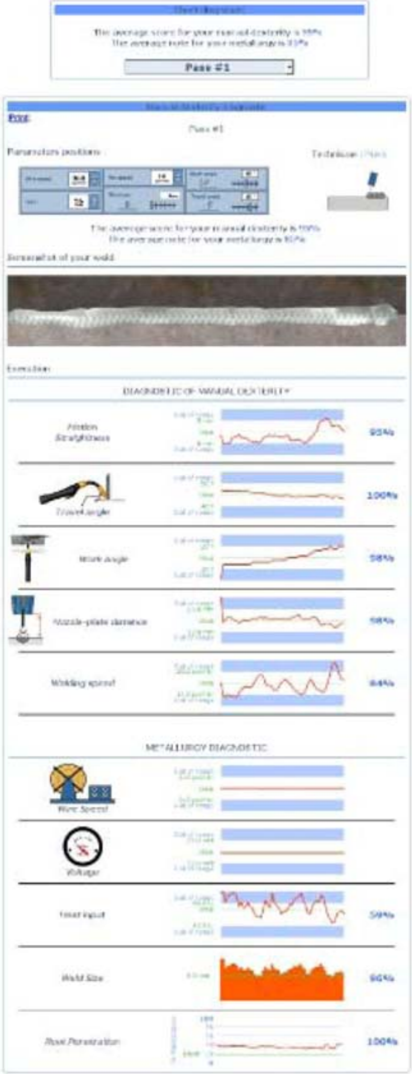


Figure 9: A weld diagnostic report gives immediate feedback, showing how close the student came to making an optimal weld. Source: 123Certification Inc.

BETA Phase Gas Tungsten Arc Welding

In the aircraft maintenance business, the quality requirements are extremely high. They demand particular care and a down-the-line exactness of the welded parts because they will be exposed to high temperatures in a combustion chamber of an aircraft engine.

Welders are regularly subjected by the quality control department to training exercises that make sure they fulfill contractual requirements.

The ARC+ simulator has been designed to meet these demands. For example, data from a specific welding code can be incorporated into a virtual weld procedure to ensure that the weld meets those criteria. Indeed the metallurgy diagnostic (second part of figure 9) shows heat input, weld size and root penetration. This has been called the eCertification® process and it allows welders or apprentices to perform virtual welds on of expensive light alloys such Titanium, Chrome-Molly or Cobalt. To develop hands-on reaction with precise welding parameters, the ARC+ simulator requires the same gesture control as with a real welding station. The welder has to replicate the same gestures, and those gestures have to be fine tuned in the same way as real qualification performance.

Open house at a Welding Training Centre in Eau-Claire, Wisconsin, USA

In March 2008, 123 Certification Inc participated at an Open House in Eau-Claire, Wisconsin, USA.

A newsletter item and pictures were prepared after that event:

"VIDEO GAME" TECHNOLOGY TO FILL GROWING NEED

A star attraction at Wisconsin's Chippewa Valley Technical College March 6 was the "ARC+" welding simulator demonstrated by its inventor, Claude Choquet of Montreal. Scores of professional welders, instructors, and people considering a welding career tried their hand at the simulator and got immediate feedback on their performance and aptitude. The occasion was the CVTC Weld Show open house, held on the school's Clairemont campus in Eau Claire, Wisconsin.

Many areas of North America are experiencing persistent shortages of qualified welders, even in a state like Wisconsin with a well-established network of technical colleges. Mr. Choquet and his company, 123Certification, have helped bolster the Quebec technical workforce with this invention and now offer it for export. The Wisconsin's CVTC event highlighted welding career opportunities in the fields of manufacturing, construction, and maintenance.

The simulator, which fits into its own suitcase for portability, uses advanced motion tracking technology and hi-resolution 3D rendering software for a virtual welding experience. Realistic simulated sounds, smoke, and sparks add to the experience. It's not all fun and games, however. When students have completed virtual welds, they can view previous welding sessions through replay mode and get advice from instructors who receive detailed reports on students' practice sessions.

There are a number of benefits the ARC+ simulator provides for welders, for industry, and for the environment. The simulator is used to train and qualify the skill levels of advanced welders and has the potential to be used in the re-certification process. Through repetitive training, welders can focus on continuous improvement of their welding skills yet reduce their exposure to fumes and sparks. The associated reduction in the use of raw materials not only provides cost saving benefits, but also reduces the impact on the environment.

Mr. Choquet, a welding engineer and expert in applying virtual reality technologies, has had an almost life-long interest in welding. His father, Professor Joseph-André Choquet, was an expert on weld fatigue at Polytechnique Montreal, a leading school of engineering [4].

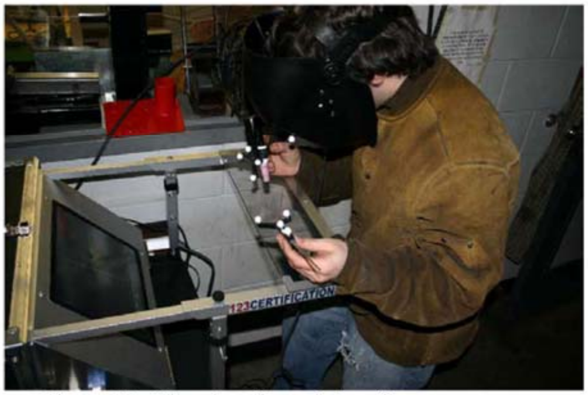


Figure 10: Virtual reality training of gas tungsten arc welding (GTAW).

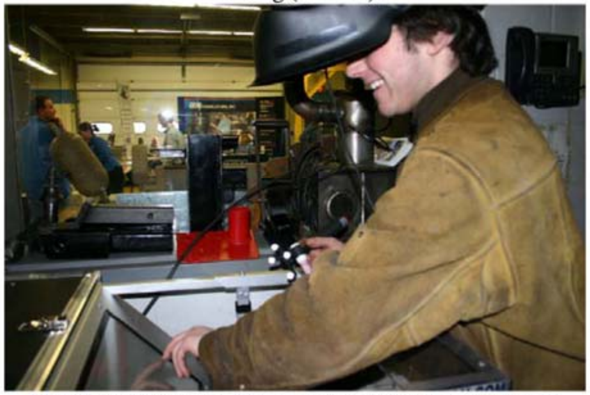


Figure 11: Tactile Screen is used to access the diagnostic allowing the welder to take immediate corrective action, and continuously improve his skills in a safe and enjoyable setting. Source: Chippewa Valley Technical College in Eau Claire, Wisconsin

3. Methodology

The Reality

We introduce the GTAW welding method with two-hand motion capture in order to be able to simulate the weld deposition. With pictures taken at a high speed, Figure 12 shows the filler metal motion activated by the first hand and the arc's motion activated by the second hand. The motion of the first hand is responsible for the solidification wave. Uniform motions of the first hand help insure metal deposit uniformity, a key criterion for compliance with aviation requirements.

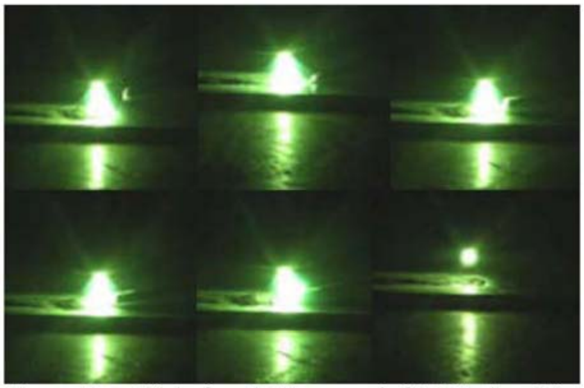


Figure 12: High Speed Pictures of GTAW deposition Source: 123Certification Inc.

Some challenges are related to the image rendering of weld pool collapsing. The work in progress is shown in the next images.

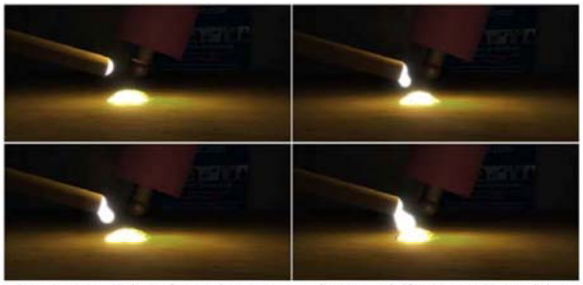


Figure 13: High Speed Images of Virtual GTAW deposition. Source: 123Certification Inc.

The motion is captured and rendered in 3D images in the glasses of the welder's mask.

Development of fine motor skills is captured, processed and posted in real time for tracking by the learner and instructor.

4. Results

School attractiveness
Since 2007, more then 150 professors at high school level introduced a free access to ARC PC in their classroom. By using a projector they could show to the students what welding is about in terms of motion dexterity with only a mouse. During the proof of concept that took place during 3 years, each students experiences at home or in a laboratory 40% of the ARC+ diagnostic mode at no costs, More then 3000 welds were performed virtually and the results shows an interest such as Gameplay Theory Principle have been showing since 20 years.

Aerospace IndustryAn aerospace industry experimenting with our technology decided to do a step by step approach. The president of this family-owned company asked his daughter - with no previous welding education - to experiment with the Arc+ simulator. This first step was intended to benchmark a beginner before training. That first step being conclusive, the QC/QA manager of that company was requested to be trained. After only four (4) 3-hours sessions during which two-thirds of the trainee's time was spent on the simulator, the QC/QA manager was able to perform a linear and uniform metal deposition as required for a welder apprentice (Fig. 14). This experiment was completed without any material consumption. The positive result led to the second step which was the test made by the quality control manager. He is in charge of producing the documents attesting to the weld's integrity. The Quality Manager had never welded before and his company proposed to him a virtual training with the Arc+ welding simulator. He later went on to complete traditional factory training.

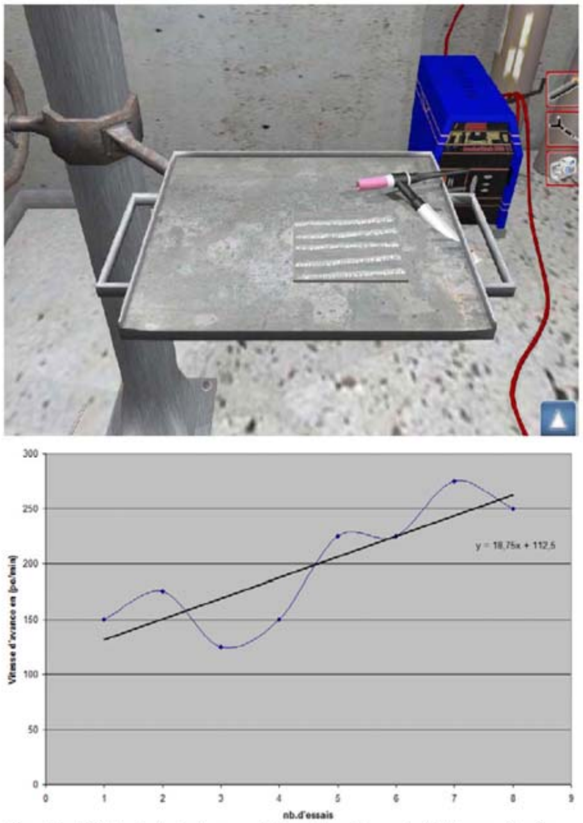


Fig 14: Virtual training welding booth and drill results from the Q4 2007 training exercise mentioned in the result section. Source: 123Certification Inc.

5. Conclusion

A key point for many training establishments is the user to simulator ratio, most training establishments wish to increase the size of the class but this would require extra training staff to supervise. ARC+ has now a brother coming from a scientific breakthrough called ARC PC. ARC PC is a internet based simulator that is an exact copy of the software that is in the Simulator itself, here the students can with the aid of an air or gyroscopic mouse train 2 of the possible 5 essential variables required by a welder to produce a good weld. This software program is offered at a ratio of software licences per simulator unit sold.. So the ARC+ system offers a much larger trainability package than any of other products.

Both products ARC+ and recently ARC PC are now fully operational for the manual and semi-automatic welding processes motion tracking along with the processing of metallurgy equations The system delivers real time 3D images with all essential variables affecting the weld pool, the arc length and the welding bead without jittering, delay and drifting. The Graphic User Interface (GUI) fully independent of any hardware technology is now available on 2 different options. The ARC+ is now a product worldwide recognized with his user-friendly, high quality graphics, multi-process, small footprint, scalability and Military precision and ruggedness. In the last year the ARC+ was delivered from North America to all corners of the world, including Australia, Germany, Kazakhstan and India.

6. Next steps in regards to mission and vision

The next step will be dictated by our clients, many of whom plan to increase the use of technologies which reduce unfavorable effects on the environment thanks to features mentioned here-after.

Environment: This is a truly green product, we normally associate a green product with what type of effect it has on our environment and this product is no exception, it is a product that is kind to our environment, it uses a minimum of energy, requires no materials, gases, consumables etc, etc etc. But what we forget is how it affects our employees, this is where this product differs from other motion tracking technology, most other technologies available on the market today use magnetic fields to detect where the work piece is in relation to the torch. To do this they must project an artificial magnetic sphere from the base unit similar to that used by mobile phones, these forms of magnetism are a hot topic at the moment due to the effect they (may) have on the individual using the unit, indeed this form of created magnetism has been decreed a pollution by several world authorities on the subject. ARC+ does not use this system, it use a passive technology with sensor located far from the user and use a form of light coming form infrared cameras that uses about the same energy as a pocket flashlight, so you can rest assured that there are no hidden dangers with our system and the training management can sleep at nights knowing they are safe.

Attraction: There is a drastic shortage in the number of young people wishing to learn the welding trade and this is not a phenomenon reserved for us Europeans/North-Americans, it is a world wide epidemic. For too long now welder education has relied on prehistoric methodology to train its students, now with simulation we can bring the training methods back into the 21 century and have a training tool that will not only attract the youth of today but also allow them to learn quicker. VR is a platform that they not only understand but also feel at home with.

Individuality: Using the Arc+ collective database the instructor can tailor make a training program for each individual trainee, catering each program to the strengths and weaknesses of the individual concerned. This allows the instructor a degree of creativity that he would never be able to achieve with the old-fashioned training methods. Each completed weld is given a diagnostic report and a score enabling the instructor (or student) to playback each weld and show the student individually or collectively where his strengths and weaknesses' lay.

Future-safe: Arc+ is a product that will develop with the technology unlike a standard welding machine. The unit is fully programmable/updateable so that when a new technology or changes to standard technologies appear on the market the unit can be updated to allow the students to take full advantage of the said changes. Hence the unit will remain longer in use than a standard welding machine and enable the students to gain all they can from changes in technology, this then produces a better class of graduate for the job market.

Hi-end: Arc+ is the perfect tool for all aspects of welder training and at all levels, we are the only system that can offer the training establishment TIG welding and are therefore the only machine that can completely encompass all hand welding processes. Arc+ in conjunction with Arc PC is also an ideal tool for adoption in the training of welding engineers, here they can learn all they need to know about the welding processes from a computer and then carry out the practical side of life on the Arc+ simulator, Arc+ will give them a diagnostic report that they will understand and be able to work with, something that is just not available in current training methods. Welding engineers need to be excited and trained using state of the art technology; Arc+ and Arc PC offer them this capability and therefore will also attract more students to this form of engineering.

Completeness: with the integration of Arc PC, Arc+ offers a complete learning package; no other simulator offers this kind of complete program for the students ensuring at all times that the group is working to achieve its goals. It introduces a level of competition so that the group dynamics also encourages the students to concentrate on their individual scoring, this in a fun way. This has been proven pedagogically to be the best way to improve the learning concentration of students.

7. Key words

Welding, Welder, Simulation, Manual, Semi-Automatic, Gesture Control, Virtual Reality, Apprentice, Professional,

8. Acknowledgements

This work was partially funded by the IRAP, NSERC and CQRDA. The Chippewa Valley Technical College in Eau Claire, Wisconsin provided their pictures from their annual Open-House activities. Our early-stage aeronautics client also helped in important ways to validate and refine our virtual training system.

References

[1] B. James, Why Welders Fail GMAW Tests, CWB Net, Sept. 1996, Vol.2, no.2

[2] T. Heston, Senior Editor, March 2008 Edition, The Fabricator Magazine, http://www.thefabricator.com/ArcWelding/ArcWelding_Article.cfm?ID=1878

[3] M. Reilly January 2008, NewScientist.com http://www.bannerhealth.com/About+Us/News+Center/In+the+News/A+Wii+warm-up+hones+surgical+skills.htm#

[4] C. Orlowek, Quebec Government Delegation at Chicago, http://www.mri.gouv.qc.ca/portail/_scripts/actualites/viewnew.asp?NewID=5516&strIdSite=chi〈 =en

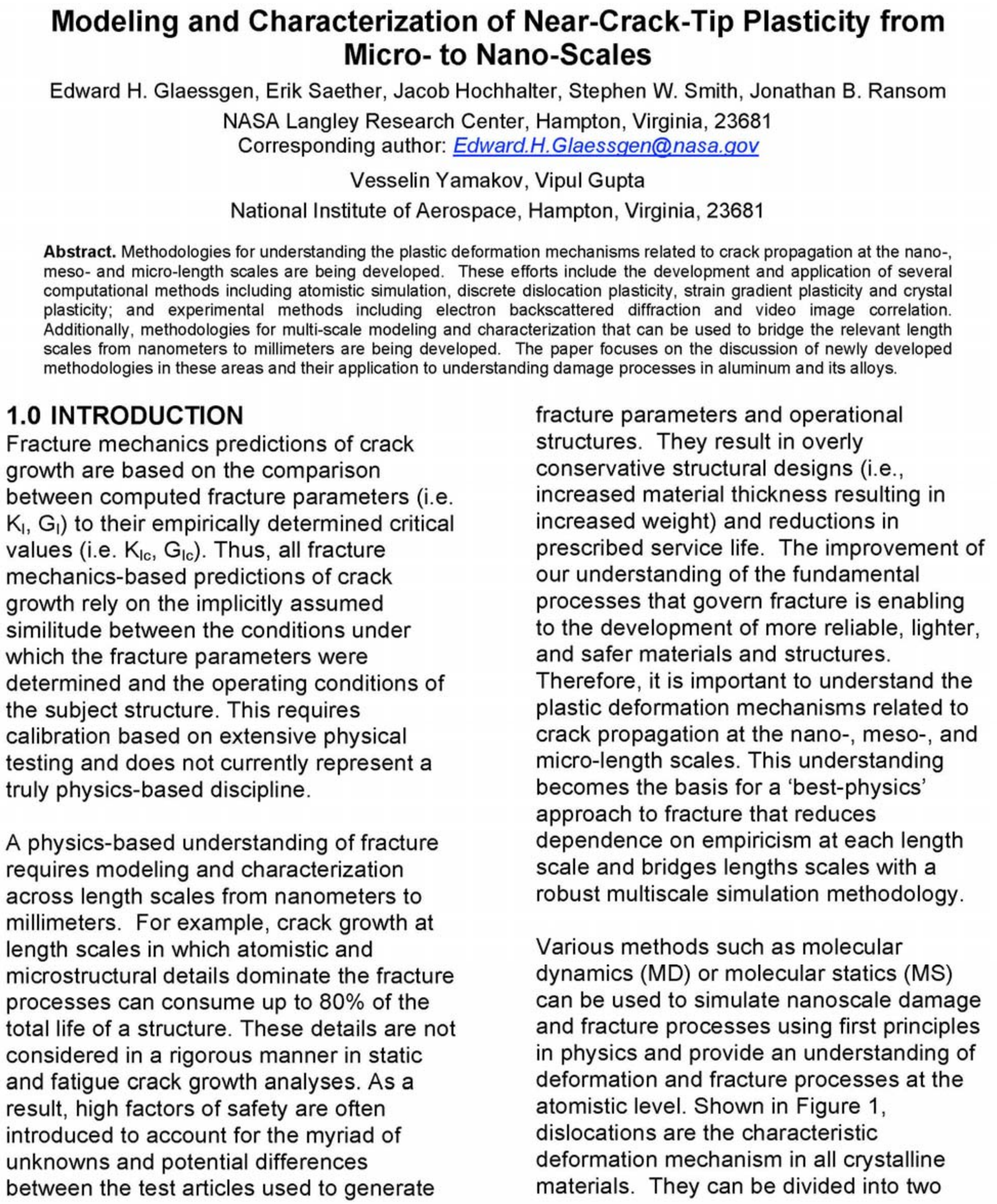
[5] P. Fuchs, G. Moreau, « Le Traité De La Réalité Virtuelle », Les Presses de l'École des Mines de Paris., Paris, 2003K. Elissa, "Title of paper," unpublished.

[6] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, B. MacIntyre. « Recent Advances in Augmented Reality », IEEE Computer Graphics & Applications, vol. 21, no. 6, 2001.

[7] J.P. Gee, « Learning by Design: Games as Learning Machine », Keynote speech, March 15 2004, RIMA 2004.

[8] 123 Certification Inc. http://www.123arc.com/en/demovideo.htm "GTAW Apprentice in action with ARC+ Welding Simulator, Q4 2007 update.

2.15 Modeling and Characterization of Near-Crack-Tip Plasticity from Micro- to Nano-Scales



Modeling and Characterization of Near-Crack-Tip Plasticity from Micro- to Nano-Scales

Edward H. Glaessgen, Erik Saether, Jacob Hochhalter, Stephen W. Smith, Jonathan B. Ransom
NASA Langley Research Center, Hampton, Virginia, 23681
Corresponding author: Edward.H.Glaessgen@nasa.gov

Vesselin Yamakov, Vipul Gupta
National Institute of Aerospace, Hampton, Virginia, 23681

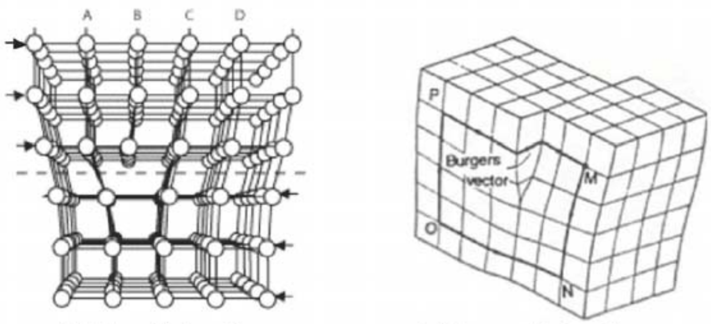
Abstract. Methodologies for understanding the plastic deformation mechanisms related to crack propagation at the nano-, meso- and micro-length scales are being developed. These efforts include the development and application of several computational methods including atomistic simulation, discrete dislocation plasticity, strain gradient plasticity and crystal plasticity; and experimental methods including electron backscattered diffraction and video image correlation. Additionally, methodologies for multi-scale modeling and characterization that can be used to bridge the relevant length scales from nanometers to millimeters are being developed. The paper focuses on the discussion of newly developed methodologies in these areas and their application to understanding damage processes in aluminum and its alloys.

1.0 INTRODUCTION

Fracture mechanics predictions of crack growth are based on the comparison between computed fracture parameters (i.e. K_I, G_I) to their empirically determined critical values (i.e. K_{Ic}, G_{Ic}). Thus, all fracture mechanics-based predictions of crack growth rely on the implicitly assumed similitude between the conditions under which the fracture parameters were determined and the operating conditions of the subject structure. This requires calibration based on extensive physical testing and does not currently represent a truly physics-based discipline.

A physics-based understanding of fracture requires modeling and characterization across length scales from nanometers to millimeters. For example, crack growth at length scales in which atomistic and microstructural details dominate the fracture processes can consume up to 80% of the total life of a structure. These details are not considered in a rigorous manner in static and fatigue crack growth analyses. As a result, high factors of safety are often introduced to account for the myriad of unknowns and potential differences between the test articles used to generate fracture parameters and operational structures. They result in overly conservative structural designs (i.e., increased material thickness resulting in increased weight) and reductions in prescribed service life. The improvement of our understanding of the fundamental processes that govern fracture is enabling to the development of more reliable, lighter, and safer materials and structures. Therefore, it is important to understand the plastic deformation mechanisms related to crack propagation at the nano-, meso-, and micro-length scales. This understanding becomes the basis for a 'best-physics' approach to fracture that reduces dependence on empiricism at each length scale and bridges lengths scales with a robust multiscale simulation methodology.

Various methods such as molecular dynamics (MD) or molecular statics (MS) can be used to simulate nanoscale damage and fracture processes using first principles in physics and provide an understanding of deformation and fracture processes at the atomistic level. Shown in Figure 1, dislocations are the characteristic deformation mechanism in all crystalline materials. They can be divided into two



a) Edge dislocation b) Screw dislocation

Figure 1. Configuration of edge and screw dislocations.

types: edge dislocations and screw dislocations as shown in Figure 1a and 1b, respectively.

Through their dynamic evolution and interaction, dislocations constitute the basis for plastic deformation and strain hardening in metallic polycrystals (see Weertman and Weertman, 1992). Because of the importance of understanding dislocation plasticity, dislocation dynamics methods (DD) have been developed that model mesoscale deformation and can provide an analytical bridge between atomistic and continuum material models.

Despite advances in MD and DD, continuum methods will continue to play a dominant role in studying the deformation and fracture of structural materials because their computational efficiency enables interrogation of large domains. Examples of relevant methods that are being developed to address plastic deformation in metallic crystals include strain gradient plasticity and crystal plasticity. These plasticity paradigms are suitable for representing deformation at the micro scale and provide a mechanistic linkage to continuum material models for analysis at structural length scales.

This paper will outline several efforts that are aimed at understanding near-crack-tip plasticity at nano-, meso-, and micro-length scales and will discuss scale-dependent damage mechanisms at each of these length scales.

2.0 MULTISCALE ANALYSIS OF NEAR-CRACK-TIP PLASTICITY

Modeling and characterization approaches for understanding crack tip plastic mechanisms at atomistic, mesoscopic and continuum length scales are presented in the following subsections. Various simulation issues such as time and length scale affects in MD, dimensionality considerations in DD, and the predictive capability of continuum analysis will also be discussed.

2.1 Modeling Near-Crack-Tip Plasticity at the Nano Scale

As a crack grows within a metallic material, a plastic zone is formed as dislocations are generated, propagated and accumulated. Atomistic-based modeling methods such as molecular dynamics provide a means of explicitly examining fundamental deformation mechanisms and have been a topic of considerable study during the past two decades (see Allen and Tildesley, 1987).

Because of its ubiquitous usage in aerospace vehicles, aluminum is of particular interest. Recently, a number of atomistic simulation studies on intergranular and transgranular crack propagation in pure aluminum have been published (Farkas, 2001; Hai, 2001; Tadmor, 2003; Yamakov,

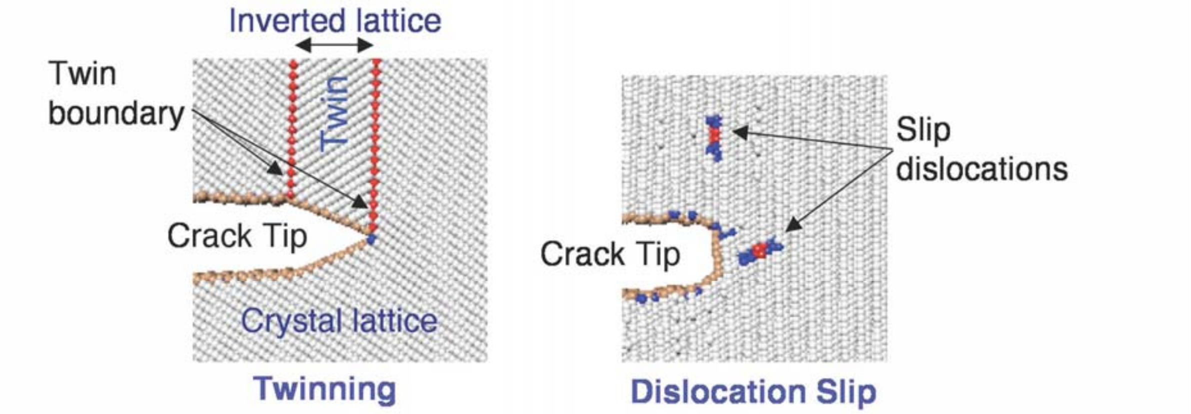


Figure 2. Twinning and slip near a crack tip in pure aluminum (Yamakov, 2009)

2006; Warner, 2007). The results of these investigations show that two main mechanisms of crack propagation and associated near-crack-tip plasticity operate at the nanoscale. These mechanisms include propagation through deformation twinning and propagation through the emission of full dislocations from the crack tip (see Figure 2). One major finding of these and other atomistic simulations of aluminum disagrees with experiment: most atomistic simulations predict deformation twinning as the dominant deformation mechanism whereas experimental observations show that dislocation slip is dominant (Tadmor, 2003).

The discrepancy between simulations and experiments has attracted considerable attention among researchers because it prevents the reliable and accurate modeling of nanoscale fracture, in particular, and puts doubt on the reliability of the atomistic simulations, in general (Tadmor, 2003; Warner and Curtin, 2007). Most likely, the source of this discrepancy is related to the very different length scales (nanometers vs. millimeters), time scales (nanoseconds vs. seconds) and stresses (GPa vs. MPa) at which simulations and experiments are usually performed. Nonetheless, the exact mechanism of how these length and time scales affect the propagation process remains unclear.

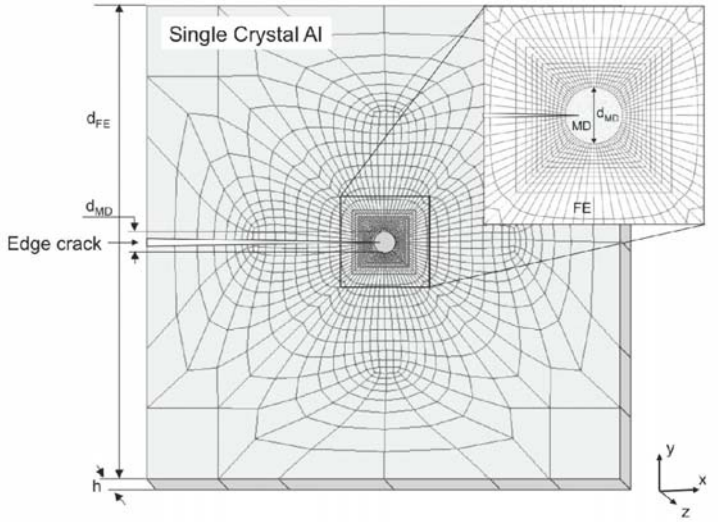


Figure 3. Model geometry of the MD-FEM coupled system with an embedded edge crack ending inside the MD domain.

To further improve the understanding of the sources of the discrepancy between simulation and experiment, a detailed study has been undertaken to determine the conditions under which twinning or dislocation emission occur at a crack tip under Mode I loading (Yamakov, 2009). The recently developed Embedded Statistical Coupling Method (ESCM; Saether, 2009) for concurrent multiscale modeling was used. ESCM was developed to enable much larger material domains to be simulated than can be considered using MD simulation alone by embedding the atomistic domain within a much larger continuum domain (see Figure 3) and thereby greatly reduce computational requirements. The atomistic region containing the crack tip is simulated using MD, while the surrounding continuum region is simulated using the finite element method (FEM). This approach was recently applied to study transgranular fracture in a single crystal of aluminum.

As shown in Figure 3, a circular MD domain was embedded in a square FEM mesh with a pre-existing edge crack propagating along the *x*-direction (Yamakov, 2009). The crack plane normal was along the *y*-direction, and the crack front was along the *z*-direction which was initially extended into the MD domain (see the enlarged central zone in the inset in Figure 3).

The crack front lies in the intersection of the $(11\bar{1})$ slip plane and the crack plane, thus the angle θ, as shown in Figure 4, is 90°. The orientation of the crack front line is chosen as the *z*-direction in the model. Under these circumstances, the mechanism of crack tip dislocation nucleation is studied as a function of the twist angle, φ, formed between the crack plane normal (the *y*-direction in Figure 4) and the [112] twin axis, lying in the $(11\bar{1})$ slip plane.

Theoretical analysis by Tadmor and Hai (Tadmor, 2003) has shown that the tendency of the crack tip to nucleate a twin or a full dislocation under mode I loading is governed by the twist angle, φ, while the tilt angle, θ, affects only the critical load of nucleation. Studying the crack tip nucleation process at a fixed angle $\theta = 90^{\circ}$, while varying the angle φ and the applied stress intensity, K_{I}, reveals the existence of a transition stress intensity, K_{IT}, below which the crack emits full dislocations and above which deformation twinning becomes dominant. A minimum value of K_{IT} is reached at $\varphi = 0^{\circ}$ where twinning becomes the dominant crack propagation mode and a maximum value of K_{IT} at $\varphi = 30^{\circ}$ defines the region of full dislocation emission at typical MD loading rates. To be consistent with experimental observations, where deformation twinning at the crack tip in aluminum is rarely observed, this study suggests that crystallographic orientations close to $\varphi = 30^{\circ}$ should be used for

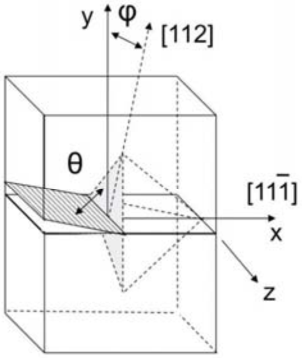


Figure 4. Crystallographic orientation of the crack with respect to the $(11\bar{1})$ slip plane in the fcc lattice.

atomistic characterization of crack tip plastic processes in aluminum. If orientations close to $\varphi = 0^o$ cannot be avoided, the results should be treated with caution as they may produce an artifact of deformation twinning.

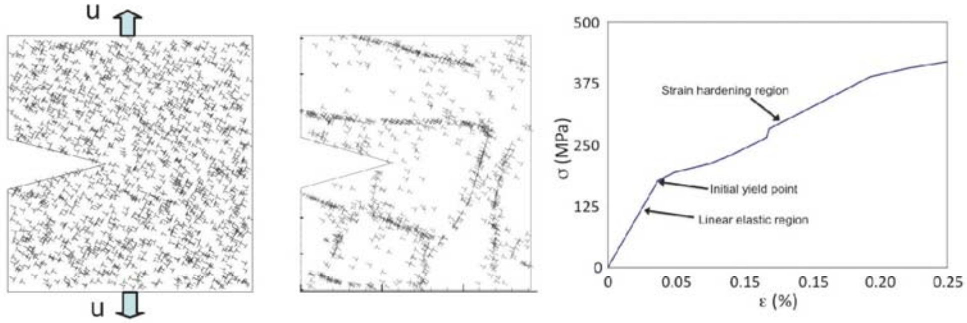
2.2 Modeling Near-Crack-Tip Plasticity at the Meso Scale

DD simulation methods have been developed to represent large numbers of dislocations, obstacles and sources discretely, but at relatively large length scales compared to atomic dimensions. In DD, the discreteness of individual atoms is homogenized with dislocations modeled as displacement discontinuities within an elastic medium and the strength of the discontinuity equal to the magnitude of the Burgers vector. Away from the core region, the displacement, stress and strain fields are suitably represented by analytic elasticity solutions. All the constitutive laws for DD are obtained directly from atomic theory, atomistic simulations, or from physical principles (Kubin et al. 1992). Simulations can involve infinite domains that are modeled using far-field boundaries or periodic boundary conditions, or as finite domains with various applied boundary conditions.

Dislocation dynamics is based on constitutive or field relations describing the short and long-range interactions between dislocations that are solved incrementally. During a simulation, the evolution of the dislocation field is obtained by forward integration of the governing equations and the plastic stress-strain relationship is directly obtained during the analysis. As discussed by van der Giessen and Needleman (1995), the computation of the deformation history is performed in an incremental manner as follows: (i) in the current state, the Peach–Koehler forces on each dislocation are determined based on the present stress fields; (ii) the change of the dislocation structure is obtained by integration of the equations of motion while governing relations are applied to test for dislocation nucleation, annihilation, short-range junction formation, and possible pinning at obstacles; (iii) updated stress state for the new dislocation configuration is repeated by returning to step (i).

Discrete dislocation plasticity simulations may be performed in either two or three dimensions. In two-dimensional simulations, dislocations are represented as point defects that are constrained to move on prescribed slip planes. This yields a simplified representation that provides qualitative results of dislocation interaction and the resulting plastic and hardening behavior of material domains. While the inelastic stress-strain and hardening behavior is an outcome of the analysis, much investigation has been made to determine the formation of dislocation structures such as subcells, shear bands, and low-angle grain boundaries.

Two-dimensional DD analysis has been enhanced by incorporating various three-dimensional processes such as junction formation resulting in dynamic sources and obstacle formation (Benzerga, 2004). An example is presented in Figure 5 in which a square aluminum domain having an edge notch is subjected to an applied external normal displacement in the *y*-direction (Figure 5a). The initial dislocation field shown in Figure 5a is randomly distributed while an evolved dislocation field showing subcell formation is shown in Figure 5b. The computed stress-strain relation is presented in Figure 5c and shows an initial linear elastic behavior followed by yielding and subsequent hardening due to dislocation interactions. Fully three-dimensional DD analyses are in the early stages of development and, although very promising, are very computationally intensive and currently have limited application (Arsenlis, 2007).



a) Initial dislocation field b) Subcell formation c) Resulting material behavior

Figure 5. Simulation using the two-dimensional DD-SIM code for a wedge configuration under normal tension loading in the *y*-direction.

2.3 Modeling Near-Crack-Tip Plasticity at the Micro Scale

Dislocations of all types are often divided into two classes in continuum mechanics-based analyses. These classes are the *statistically stored dislocations* (SSDs) that are generated by the manufacturing processes used to form the material (pre-existing dislocations) and by uniform deformation during loading, and the *geometrically necessary dislocations* (GNDs) that are required to enforce internal compatibility during non-uniform plastic deformation.

Both SSDs and GNDs are explicitly considered in molecular dynamics and dislocation dynamics models; however, neither is explicitly considered within conventional crystal plasticity (CCP) formulations. CCP formulations, like all constitutive models, must be calibrated for the specific material of interest. In CCP, the calibration usually consists of modeling a material microstructure with specific grain size, aspect ratio and crystallographic orientation, and matching a simulated response to an observed response by varying the material parameters.

Unlike CCP theories, the basic tenant of strain gradient plasticity (SGP) theories is that GNDs are produced by micron-scale gradients at a density comparable to, or greater than, that of SSDs, thus increasing the total dislocation density and the resistance to plastic flow. It is now generally accepted that any apparent increase in flow strength is due to the generation and storage of GNDs as required to maintain internal compatibility during non-uniform plastic deformation, e.g., localized gradients near crack tips or precipitates. Numerous SGP theories have been proposed recently with the purpose of extending the validity of continuum plasticity theories down to the micron scale. Inherent to SGP is the presence of a characteristic length scale over which the underlying mechanisms of plastic response are dependent on the magnitude of strain gradients. In addition to providing a more accurate model for work hardening, strain gradient plasticity models are closely related to and may be calibrated by the results of dislocation dynamics simulations, thereby providing a more natural tie to inelastic deformation processes at lower-length scales.

These microscale plasticity models can be integrated with detailed three-dimensional finite element models of microstructure to provide a new understanding of microstructural deformation. By incorporating models for slip accumulation, a relationship between plastic exhaustion and crack growth can be computed. Detailed three-dimensional finite element models of aluminum 7075-T651 employing

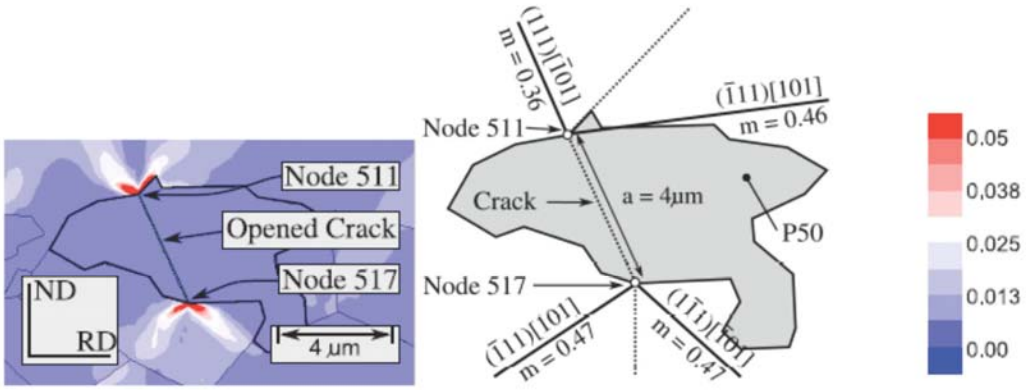


Figure 6. Computed slip fields near a cracked constituent particle that was observed to nucleate a crack into the surrounding grains.

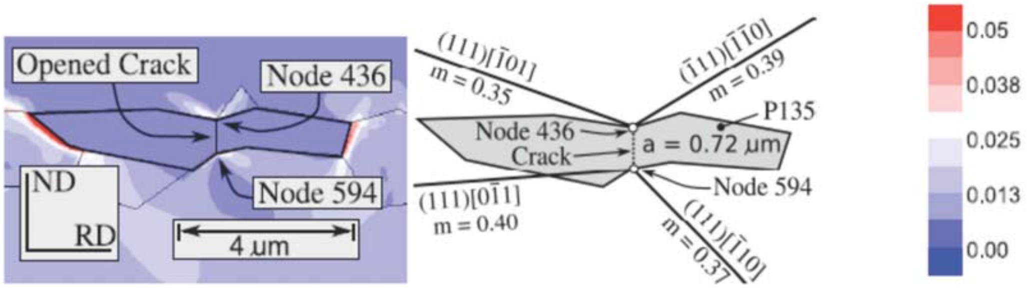


Figure 7. Computed slip fields near a cracked constituent particle that was observed not to nucleate a crack into the surrounding grains.

CCP have been generated to predict the initiation of cracking at the microstructural scale (Hochhalter, 2010). To better understand the slip accumulation during cyclic loading that precedes nucleation events, these finite element models were generated using observed microstructural data that included the configuration and location of constituent particles and grain microstructural details. Slip localization and accumulation was computed near cracked particles. Figures 6 and 7 illustrate the computed slip localization near two different cracked constituent particles in aluminum 7075-T651. The contoured fields in both figures are the maximum value of slip on any one of the twelve fcc slip systems; the corresponding values given by the contour bars are the magnitude of slip on the dominant system.

The particle shown in Figure 6 with high slip accumulation near the crack tip-grain interface was observed to nucleate a crack into the surrounding grains, while the particle in Figure 7 did not. Thus, it appears that slip localization and accumulation plays a governing role in crack nucleation at this scale; see Hochhalter (2010) for further discussion. Figures 6 and 7 also show the correspondence between computed slip localization and dominant slip system directions as measured using electron backscattered diffraction (EBSD). However, the directions of slip localization did not correspond with the nucleation direction given by the dotted line in Figure 6. This observation leads to two possible hypotheses, that crack trajectories are based on alternating shear or on local maximum tangential stress. More simulations and experimental characterizations are currently underway to investigate these hypotheses.

2.4 Characterization at the Meso and Micro Scales

Even though considerable progress in computational methodologies and algorithms has been made, an in-depth understanding of damage processes is still heavily dependent on experimental characterization. Small-scale experimental methods are being developed to understand damage processes and validate simulations at the meso and micro length scales. In one example, an experimental methodology that uses an environmental scanning electron microscope (ESEM) equipped with *in situ* loading frame and EBSD system has been employed to characterize damage processes in single crystals of pure aluminum and polycrystalline aluminum alloys. The EBSD orientation mapping tools can be used to measure the extent of high plastic deformation near the fatigue crack tip and crack tip wake. Plasticity near the crack tip is related to the plastic strain gradients and thus the geometrically necessary dislocation (GND) density.

These GNDs result in bending of the lattice and may be detected as an orientation gradient within a single grain. Additionally, a zone of "significant plastic strain" about a fatigue crack tip and crack tip wake can be determined by measuring the width of the highly defected region (e.g., green-to-red rainbow color scheme on misorientation maps). Experimentally determined locations of orientation discontinuities, e.g., at sector boundaries, slip bands, near the crack-tip and GND densities estimated from local lattice rotations can be compared with model predictions to enable the *physics-based* models to include correct input parameters, such as source and obstacle densities.

Recent studies (Sun, 2000; Kysar, 2002) of single crystals and bicrystals have shown that it is possible to extract some of the components of the Nye dislocation density tensor (Nye, 1953) using orientation data obtained by EBSD mapping, provided that the crystal orientation and deformation conditions are carefully controlled to constrain the number of independent components. The present work follows that of Sun (Sun, 2000) and Kysar (Kysar, 2002), and considers a connection between the GND content and the lattice curvature tensor through spatially resolved local-orientation measurements using EBSD.

For the purpose of illustration, these approaches have been applied to the EBSD orientation data obtained from the vicinity of a fatigue crack in precipitation-hardened aluminum alloy Al-Cu-Mg 2024-T351. The intra-grain misorientation map (Figure 8a) displays changes in the local orientation, along with large amounts of intra-granular misorientation associated with the large plastic deformation in the vicinity of a crack tip wake (Gupta, 2009). White regions in Figure 8a correspond to pixels that were not indexed. The misorientation map reveals distinctions in the morphology of plastic damage, e.g., the presence of slip-bands near the crack tip wake. These maps suggest the presence of a high dislocation content resulting in extensive misorientation.

Figure 8b shows the estimated distribution of GND density within the scanned area. The regions of lower dislocation density (i.e., base material, ~0.5-1x10^{14}/m^2) are separated by regions of higher dislocation density (i.e., *plastically-deformed* crack-wake, ≥10^{15}/m^2 and higher), and can be identified by marked orientation change (Figure 8a) or by the enhanced dislocation density (Figure 8b; Gupta, 2009). The boundaries of these banded structures (dislocation patterning) contain a high GND density, and regions within the bands are relatively free of dislocations that contribute to lattice curvature. An inhomogeneous distribution of the dislocation density becomes obvious for such cases.

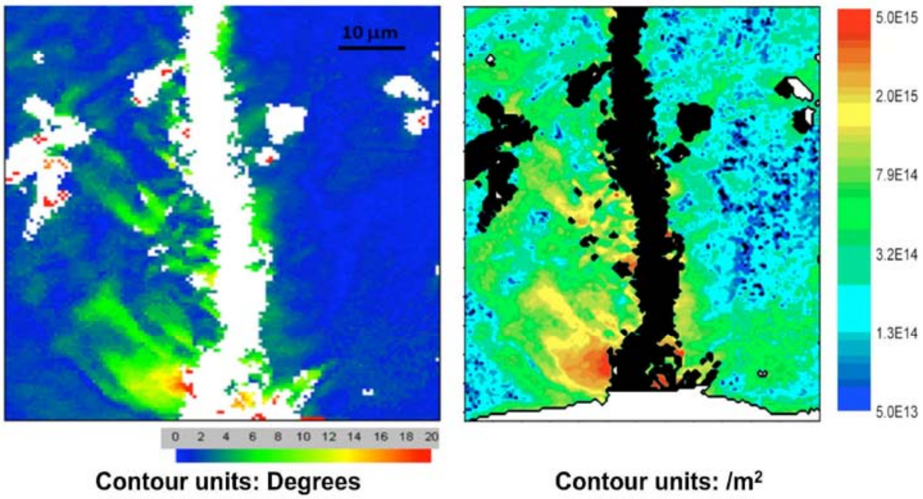
The measurements of local orientation changes and estimates of GND content near the crack tips and deformed wakes of fatigue cracks can be qualitatively compared with those predicted by computational models developed with the aid of molecular dynamics and finite element simulations. This experimental effort will contribute a significant quantitative and physical understanding of damage mechanisms that will enable next-generation damage models to progress beyond the current empirical models.

3.0 SUMMARY

An examination of plastic mechanisms as a function of length scale reveals that the phenomena governing plasticity become increasingly complex as the length scales increase. At the nanoscale, plasticity is characterized by the formation and movement of individual dislocations. Assumptions at the nanoscale are related to very fundamental quantities such as the interatomic potential, structure of the crystal lattice, and the presence of alloying elements. At the mesoscale, the discrete dislocations interact with each other, with sources and with obstacles. DD simulations can model the behavior of micron-sized domains, but must use aggregated values of source and obstacle strength and spacing in addition to approximate solutions for dislocation interaction. At the microscale, the complex physical interactions within a grain interact with those occurring in neighboring grains and form an even more complex deformation field. Crystal plasticity and strain gradient plasticity formulations can account for plastic slip in a homogenized sense, but must be calibrated against experimental data or smaller-scale simulations.

Thus, a consequence of the limitations of existing modeling tools and finite computer resources is that the simulation of increasingly complex domains necessitates a corresponding decrease in the fidelity of the analyses.

The promise of these various computational methods for modeling near-crack-tip plasticity is not seen when the methods are implemented individually, but rather, when the methods are integrated with each other and with the understanding gained from small-scale experiment. For example, molecular dynamics simulations can be used as the basis for discrete dislocation plasticity models, which, in turn, can be used to inform the parameters needed in gradient and crystal plasticity models. The results from small-scale experiment can then be used to validate and augment these simulations.



(a) EBSD misorientation map (b) Enhanced dislocation density map

Figure 8. Maps of misorientation and geometrically necessary dislocation density

4.0 REFERENCES

Allen, M.P. and Tildesley D.J., *Computer Simulation of Liquids*, Oxford Science Publications, Oxford, 1987.

Arsenlis, A., Cai, W., Tang, M., Rhee, M., Oppelstrup, T., Hommes, G., Pierce, T.G. and Buylatov, V.V., "Enabling Strain Hardening Simulations with Dislocation Dynamics," *Modeling and Simulation in Materials Science and Engineering*, vol. 15, 2007, pp. 553-595.

Benzerga, A., Brechet, Y., Needleman, A. and Van der Giessen, E., "Incorporating Three-Dimensional Mechanisms into Two-Dimensional Dislocation Dynamics," *Modeling and Simulation in Materials Science and Engineering,* vol. 12, 2004, pp. 159-196.

Farkas, D., Duranduru, M., Curtin, W. A., Ribbens, C., "Multiple-Dislocation Emission from the Crack Tip in the Ductile Fracture of Al," *Philosophical Magazine A*, vol. 81, 2001, pp. 1241-1255.

Gupta, V.K., Ph.D. Dissertation, University of Virginia, Charlottesville, 2009.

Hai, S. and Tadmor, E. B., "Deformation Twinning at Aluminum Crack Tips," *Acta Materialia.*, vol. 51, 2003, pp. 117-131.

Hochhalter, J.D., Littlewood, D.J., Christ Jr., R.J., Veilleux, M.G., Bozek, J.E., Ingraffea, A.R., Maniatty, A.M., "A Geometric Approach to Modeling Microstructurally Small Fatigue Crack Formation: II. Physically based modeling of microstructure-dependent slip localization and actuation of the crack nucleation mechanism in AA 7075-T651," *Modeling and Simulation in Materials Science and Engineering*, vol. 18, 2010, pp. 1-32.

Kubin, L.P., Canova, G., Condat, M., Devincre, B., Pontikis, V. and Brechet, Y., "Dislocation Microstructures and Plastic Flow: A 3D Simulation," *Solid State Phenom.*, vol. 23/24, 1992, pp. 455-472.

Kysar, J.W. and Briant, C.L., "Crack Tip Deformation Fields in Ductile Single Crystals," *Acta Materialia*, Vol. 50, 2002, pp. 2367-2380.

Nye, J.F., "Some Geometrical Relations in Dislocated Crystals," *Acta Metallurgica*, Vol. 1, 1953, pp. 153-162.

Saether, E., Yamakov, V. and Glaessgen, E.H., "An Embedded Statistical Method for Coupling Molecular Dynamics and Finite Element Analyses," *International Journal for Numerical Methods in Engineering*, vol. 78, 2009, pp. 1292-1319.

Sun, S., Adams, B.L. and King, W.E., "Observations of Lattice Curvature Near the Interface of a Deformed Aluminum Crystal," *Philosophical Magazine A*, Vol. 80, No. 1, 2000, pp. 9-25.

Tadmor, E.B. and Hai, S., "A Peierls Criterion for the Onset of Deformation Twinning at a Crack Tip," *Journal of the Mechanics and Physics of Solids*, vol. 51, 2003, pp. 765-793.

Van der Giessen, E. and Needleman, A., "Discrete Dislocation Plasticity: A Simple Planar Model," *Modeling and Simulation in Materials Science and Engineering,* vol. 3, 1995, pp. 689-735.

Warner, D.H., Curtin, W.A. and Qu, S., "Rate dependence of Crack-Tip Processes Predicts Twinning Trends in f.c.c. Metals," *Nature Materials*, vol. 6, 2007, pp. 876-880.

Weertman, J, Weertman, J.R., *Elementary Dislocation Theory,* Oxford University Press, Inc., New York, New York, 1992.

Yamakov, V., Saether, E., Phillips, D.R. and Glaessgen, E.H., "Molecular-Dynamics Simulation-Based Cohesive Zone Representation of Intergranular Fracture Processes in Aluminum," *Journal of the Mechanics and Physics of Solids*, vol. 54, 2006, pp. 1899-1928.

Yamakov, V., Saether, E., and Glaessgen, E.H., "A Continuum-Atomistic Analysis of Transgranular Crack Propagation in Aluminum," 50th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference and Exhibit, Palm Springs, CA, May 4-7, 2009.

2.16 Markov Chain Monte Carlo Bayesian Learning for Neural Networks

Markov Chain Monte Carlo Bayesian Learning for Neural Networks

Michael S. Goodrich
Old Dominion University
mgood028@odu.edu

Abstract. Conventional training methods for neural networks involve starting at a random location in the solution space of the network weights, navigating an error hyper surface to reach a minimum, and sometime stochastic based techniques (e.g., genetic algorithms) to avoid entrapment in a local minimum. It is further typically necessary to preprocess the data (e.g., normalization) to keep the training algorithm on course. Conversely, Bayesian based learning is an epistemological approach concerned with formally updating the plausibility of competing candidate hypotheses thereby obtaining a posterior distribution for the network weights conditioned on the available data and a prior distribution. In this paper, we developed a powerful methodology for estimating the full residual uncertainty in network weights and therefore network predictions by using a modified Jeffery's prior combined with a Metropolis Markov Chain Monte Carlo method.

1.0 INTRODUCTION

We propose a methodology for estimating the full residual uncertainty in Artificial Neural Network (ANN) weights and therefore network predictions by using Bayesian probability analysis[4] (BPA), and a modified Jeffery's prior combined with computational sampling methods including Markov Chain Monte Carlo.

In this paper we restrict our attention to three layer feed-forward perceptrons, since they are sufficient[1,2] to serve as universal approximating functions. We further restrict attention to *supervised learning*. We will also not be considering feature extraction. As this effort is concerned with digital simulation based approaches, we will be using numerically driven discrete formulations (i.e. sums instead of integrals) throughout.

Artificial neural networks

An Artificial Neural Network can be thought of as a computational model which consists of three layers of processing units with full interconnection between layers such that each component of an input vector is scaled individually for each middle layer unit, and the scaled components are then summed and passed through a transfer or *activation* function in each unit in the middle layer and then the middle layer outputs constitute another vector as the input to the output layer which is likewise scaled independently and individually for each output unit. The units in the output layer are typically (but not necessarily) simple linear functions. The input vectors for each layer also contain an implicit component of 1 to serve as an input *bias*. Figure 1 depicts a conventional three-layer feed forward perceptron network.

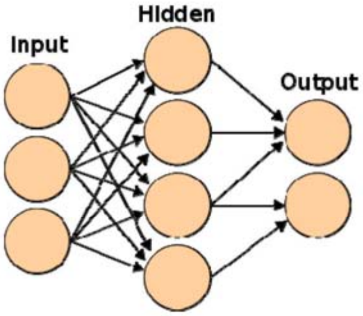


Figure 1 - Three Layer Feed Forward Network

Expressed as a mathematical model for the simplest case of a one input, one hidden unit, and one output we can write

$$y = v_0 + v_1\psi(w_0 + w_1 x) \qquad (1)$$

where *y* is the output, $\vec{w}$ are the weights (scale factors) from the inputs $\vec{x}$ (which includes an implicit bias input of 1) to the hidden layer, $\vec{v}$ are the weights from the transfer function outputs (which includes an implicit bias input of 1) to the output layer, and ψ is the non-linear activation function. Note the location of the bias components for v_0, and w_0.

For arbitrary numbers of inputs, hidden units, and outputs, equation (1) takes the form

$$y_k = v_{0k} + \sum_{h=1}^{Nh} v_{hk}\psi\left(w_{0h} + \sum_{i=1}^{Ni} w_{ih}x_i \right) \qquad (2)$$

and can be written as a matrix formulation.

.

Bayesian Probability

BPA requires determining a Universe of Discourse (UOD) which is a set of hypotheses that are ranked on a common scale of [0,1] in terms of their relative strength as an explanator of the observed data. This is done both for a family of competing models, and for competing sets of the parameter values for each model. The basic process is:

- determine a prior distribution for the model parameters of a given model
- determine a probabilistic likelihood function for the phenomena under study
- determine a UOD for our analysis
- determine a posterior distribution for the hypotheses in the UOD
- make inferences from the posterior yielding full accounting for the residual uncertainty of the parameters

This probability is then interpreted as a measure or weighting of the amount of inferential support[3] from the observed data for the hypotheses normed to entail the chosen UOD.

We wish to stress that the choices of prior, likelihood, and UOD represent *degrees of freedom* for the researcher; BPA only promises to give us the most logically justifiable results contingent on these choices[3,4].

Likelihood Model

The likelihood function is entirely dependent on the phenomena under study and must be constructed to yield the conditional probability of any observed data for a chosen model and values of its parameters.

Bayesian Prior Selection

Bayesian prior selection is a vast subject. Typically one may express ignorance concerning the current problem, or may possess some information that may be codified into a prior, e.g. by the Principle of Maximum Entropy[4,10,11]. In any case, the prior expresses our starting information concerning the parameters of the likelihood function.

Posterior Distribution

Bayesian posterior determination proceeds by computing Bayes Rule[4,13]

$$P(\vec{w} \mid D,M) = \frac{P_0(\vec{w} \mid M)L(D \mid \vec{w},M)}{\sum_{\{\vec{w}\}} P_0(\vec{w} \mid M)L(D \mid \vec{w},M)}$$

(3)

where $P(\vec{w} \mid D,M)$ is the posterior probability of the model parameters $\vec{w}$ conditioned on the observed data D and the choice of model M, $P_0(\vec{w} \mid M)$ denotes the *prior* probability of the parameters $\vec{w}$ which summarizes all knowledge of $\vec{w}$ for this model brought forward into the present analysis, and $L(D \mid \vec{w},M)$ is the likelihood of the data being observed for the model given that the parameters have the value $\vec{w}$. The denominator of (3) is known as the *evidence* for the model:

$$P(D \mid M) = \sum_{\{\vec{w}\}} P_0(\vec{w} \mid M)L(D \mid \vec{w},M) \quad (4)$$

and is the probability of the data given the model marginalized by the likelihood function over the hypothesis space $\{\vec{w}\}$.

Parameter Estimation

Bayesian parameter estimation proceeds by evaluating and scoring on a common probability scale, each value of the parameters in the UOD of model parameters using (3), thereby producing a normalized probability distribution (or mass function) over the UOD.

Model Selection

From (4) and using Bayes Rule (3) we can write

$$P(M \mid D) = \frac{P_0(M)P(D \mid M)}{\sum_{\{M'\}} P_0(M')P(D \mid M')} \qquad (5)$$

This posterior distribution over a separate UOD {M} *for models* allows selection of the model which best explains the observed data, again also yielding a full characterization of the residual uncertainty conditioned upon a choice of prior for the models and available data.

Occam Factors

BPA has an interesting feature where model selection is concerned in that it contains an explicit built in penalty for more complex models over simpler models. This feature is known as an Occam factor[11] and is a consequence of forming the ratio of the evidence for two competing models, one of more complexity than the other using equation (4). A factor that emerges in the ratio calculation will penalize[11] the more complex model due to its' greater expanse of parameter space that will be ultimately ruled out by conditioning on the available data.

Bayesian Predictions

Bayesian inference or prediction is generally concerned with the formal marginalization over the hypothesis space $\{\vec{w}\}$ of a given model. For example we might wish to check the probability of some desired output data $\vec{t}$ conditioned on our model, and our training data such that

$$P(\vec{t} \mid M, D) = \sum_{\{\vec{w}\}} P(\vec{t}, \vec{w} \mid M, D)$$

$$= \sum_{\{\vec{w}\}} P(\vec{t} \mid \vec{w}, M, D) P_0(\vec{w} \mid M) \qquad (6)$$

using the *product rule* of probability[4]. This is the predictive distribution for observing the data $\vec{t}$ formally treating the models parameters as *nuisance parameters*. For example if $P(\vec{t} \mid M, D)$ is materially different than the likelihood we might suspect our choice of likelihood function.

We might also form a simple *expectation* such that:

$$\langle \vec{y} \rangle = \sum_{\{\vec{w}\}} \vec{y}(\vec{w}) P(\vec{w} \mid D, M) \qquad (7)$$

where $P(\vec{w} \mid D, M)$ is given by (3), $\vec{y}(\vec{w})$ could be the output of an ANN with parameters $\vec{w}$, and the expectation is conditioned on the training data set D and choice of network represented by M.

Learning for Neural Networks

There remains the issue of determining the values of the weights $\vec{w}$ and $\vec{v}$ in (2) conditioned on the available data and any other relevant information. This is the central problem of ANN *learning*. We would like to point out that it is possible to determine the vector $\vec{v}$ as function of $\vec{y}$ and $\vec{w}$ with appropriate mathematical technique.

Backpropagation

The most common conventional (non Bayesian) approach to ANN learning is to concern oneself with an error function such as:

$$E = \frac{1}{N_p} \sum_{k=1}^{N_k} \sum_{p=1}^{N_p} \left[t_{p,k} - y_{p,k} \right]^2 \qquad (8)$$

As written, this is the mean squared error per pattern average for all outputs where N_k is the number of outputs, $t_{p,k}$ denotes the k^{th} desired output for the *p*th input pattern, $y_{p,k}$ denotes the k^{th} observed output for the p^{th}

input pattern, and N_p denotes the total number of observations or training data patterns.

The basic stratagem of backpropagation is to substitute (2) for the $y_{p,k}$ in (8), and then use error gradient information for the weights in order to use a numerical error reduction algorithm (typically *conjugate gradient*) to adjust the weights and achieve some error minimum.

Two outstanding issues emerge in this approach to learning:

- what is the proper minimum for the residual error that the network produces? This is the issue of *regularization* which is inhibition of network training to the noise component of the signal.

- What network model best explains the observed data?

Bayesian Learning for ANNs

In broad strokes, Bayesian Learning requires choosing a prior distribution over the network weights, framing a probabilistic formulation for an ANN model or models then using (2), (3), and (4) to determine the best network along with a posterior probability distribution over the weights for the selected model with a full characterization of the residual uncertainty in both. We describe our solution to these in section 2.

Monte Carlo Simulation

Because the Bayesian posterior - which is the sought after distribution - is a priori unknown, we must resort to some form of search strategy to find it. Monte Carlo simulations were originally developed to provide numerical integration of functions but can be used in a variety of ways to sample the solution space and determine the probability distribution for our chosen UOD, which we are choosing probabilistically as a Markov Chain rather than deterministically.

2.0 METHODOLGY

The principle challenge of our methodology is to combine Bayesian Probability, mathematical models of ANNs, and simulation based methods of solution search to determine a joint posterior probability distribution for the hidden network weights and any other parameters such as the noise or stochastic contribution to the observed data. Thus is created all that is necessary to make predictions with full accounting of the residual uncertainty in the inferred network.

Probabilistic Likelihood Model

We must determine a likelihood model to be used in equations (3),(4)

To address the issues of regularization (over fitting inhibition) and to account for actual residual stochasticity in the data, we choose to compose our “meta-model” as a linear combination of deterministic and non-deterministic or *stochastic* components. This requires expanding our hypothesis space to also ascertain the correct amount of stochasticity or loosely *noise* in the input data. To clarify, we seek to model the residual stochastic (loosely noise) component of the input data or signal as a form of regularization. To that end we model the likelihood of the target data (training pattern) less the output of the candidate network output $\vec{y}$ as:

$$L(\vec{t}-\vec{y}\mid\vec{w},\Sigma)=G(|\vec{t}-\vec{y}(\vec{w})|,\Sigma) \qquad (9)$$

where *G(.)* is a Multivariate Gaussian probability density, $\vec{t}$ is the training pattern output data, and the components of $\vec{y}(\vec{w})$ are given by (2). That is to say, our likelihood model for the difference between the target (training) data and the candidate network output is to be modeled as a form of *Multivariate Gaussian white noise*. Note that since white noise is uncorrelated, our likelihood model is *conditionally independent* between training inputs; we therefore model the stochastic part of the output as conditionally independent between training patterns while allowing for the possibility of correlations between the components of the output vector. We therefore write for the likelihood of the training set for a given choice of model parameters $\{\vec{w}, \Sigma\}$:

$$L(\vec{t}-\vec{y} \mid \vec{w}, \Sigma)=\prod_{p=1}^{N_p} G(\vec{t}_p-\vec{y}_p, \Sigma) \qquad (10)$$

Note that we have expanded our parameter search according to Bayesian principles to include the multivariate noise contribution Σ, the full covariance matrix for the difference between the observed output and modeled output. The full covariance matrix provides for possible correlations among the elements of the vector $\vec{t}_p-\vec{y}_p$ which is the modeled stochastic component associated with each training pattern vector $\vec{t}_p$. We use equation (10) for the likelihood function $L(\bullet)$ in (3) where the data D is now understood to be stochastic part of the training data i.e., $\vec{t}-\vec{y}(\vec{w})$ thus we are *absorbing* the deterministic part of the signal into the network output in such a manner as to maximize the probability associated with the stochastic *residue* of the training input via our likelihood function.

Using ANNs in this fashion can be thought of as a form of Bayesian *non parametric* probabilistic modeling with the choice of activation function serving as the appropriate basis functions[10]. (N.B.: the term "Non-Parametric Bayesian" has acquired a different meaning in the literature than what we are implying in this study)

Choice of Prior Distribution

Our methodology addresses the choice of a prior distribution by choosing a (modified) *Jeffreys'* prior[5,6,11] to express partial ignorance over the parameter space but also because it discriminates against excessively large model parameter values[12] (called *shrinkage* in the statistics literature, and *weight decay* in the ANN literature).

Jeffreys' prior is invariant to transformations of the parameter space and is related to the expected value of the Fisher Information Matrix. For scale parameters, this becomes

$$P(w \mid M)=c_0 / w \qquad (11)$$

where c_0 is a normalizing constant. This represents a density which is apportioned equally per decade of its scale and is therefore scale invariant. While the continuous version of this density is strictly improper (the cumulative distribution integral diverges), it is straightforward to construct a normalized discrete probability mass function over some chosen (always finite) UOD.

We consider the sought after network weights $\vec{w}$, and the noise contribution Σ to both be scale parameters. We modify the Jefferys' priors for both according to the following considerations:

- We impose a minimum value for each component of $\vec{w}$, and the diagonal components of Σ such that any values less than these cutoffs decay smoothly to zero
- We normalize the resulting discrete distributions over some reasonable range
- Since weight parameters $\vec{w}$ may be negative, we actually use the absolute value in (11), keeping the distribution in that case symmetric about the origin.

The resulting distributions have the general form of Figure 2 below.

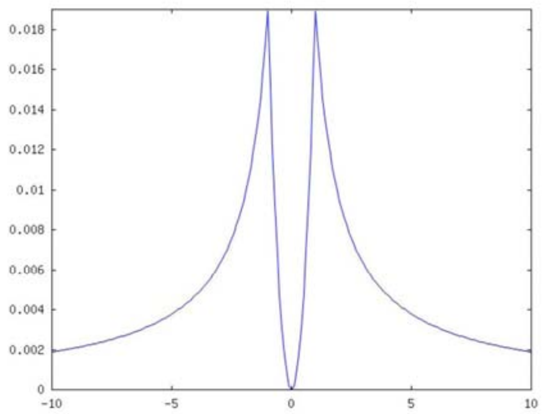


Figure 2 – Modified Jeffery's' prior density

For the diagonal components of the noise contribution parameter Σ we have only the positive (properly normalized) part of the curve. We do not discriminate against small values of the off-diagonal (covariance) elements of Σ

Strictly speaking, the cutoffs are somewhat arbitrary and good candidates for a hyperparameterized prior for each (in contemporary Bayesian fashion), but that is not included in this analysis. The actual choices made were such that the cutoff value were chosen sufficiently small to be hopefully good for a wide choice of problems.

These minimum are thought reasonable on the basis that network weights which are too small lead to uninteresting solutions, and if the noise contribution is too small then we are in effect eliminating that component of the modeling. In both cases, parameter values of 0 are clearly uninteresting.

MCMC

Choosing a UOD by Monte Carlo (MC) simulation tends to take one of two basic tracks:

1. Start at a random location and use a local rule to choose the next location
2. Use a global rule to choose the all locations

Conventional MCMC sampling techniques such as the Metropolis, Metropolis-Hastings, and Gibbs sampling are basically of type 1.

Independence Chain sampling, and Importance sampling are of type 2.

In our version of *grid or mesh sampling* we used a coarse grid to fine grid progression to characterize the posterior distribution and locate promising regions which were subsequently explored with a finer grid. This approach suffers from exponential increase in computational effort with increased dimensionality of the parameter space. The basic approach is to compute equation (3) for each grid point in the UOD, thus achieving a coarse grained posterior probability distribution. Finer grained computations over more promising regions then ensued. It is in this sense that a non-local or global rule is used in choosing sample points.

In a *random walk* oriented MCMC approach we used a Metropolis algorithm which generally is able to find promising regions, is less computationally demanding, but often may not give a complete characterization of the posterior distribution and in can yield lower quality network output when compared against the training data than grid sampling. The Metropolis algorithm operates by choosing its next point by constructing a Markov Chain via sampling from a *local proposal density* which is centered on the current point and for our study is a Multivariate Gaussian of the same dimension as the hypothesis space. The new point is then accepted with probability

$$P = \min\left\{1, \frac{P(\vec{w}_{new} \mid D, M)}{P(\vec{w}_{current} \mid D, M)}\right\} \qquad (12)$$

A chain of N points $\{\vec{w}\}$ is thus determined from this algorithm such that if a new candidate point is rejected, a copy of the current point is added to the chain. In this fashion, points are accumulated according to their relative probability, the duplication of points serving to increase their respective weighting for inferences from the chain such as expectations according to

$$\langle \vec{y} \rangle = \frac{1}{N} \sum_{k=1}^{N} \vec{y}(\vec{w}_k) \qquad (13)$$

where $\vec{y}(\vec{w})$ is given from (2), and the inferential locus is the chain $\{\vec{w}\}$.

Thus the UOD in our methodology is determined stochastically by the sampling algorithm. Each sampled point accepted or rejected may be retained so as provide for a proper discrete probability distribution of the form

$$PD = \{\{\vec{w}_1, p_1\}, ..., \{\vec{w}_K, p_K\}\} \qquad (14)$$

for a distribution with K elements to be used in equations (6), and (7). Conventional MCMC doctrine uses (13). The proposal density used in (12) must be tuned in order to achieve acceptance rates of between 25% and 50% as recommended by the literature[14]. Typically MCMC sampling includes a *warm up time* to allow the chain to begin properly representing the target posterior distribution and so the warm up time is not included in equation (13).

3.0 EXPERIMENTS

General Remarks

We performed experiments on both simulation generated data and real world data. Simulation generated data is especially beneficial for validation of the methodology since the noise uncontaminated input data is readily available. Naturally, performing well in such a case gives one confidence in attacking real-world problems where the noise component may be unknown.

Noisy Sine Wave

This experiment is for a noisy sine wave and was pattern matched with a network consisting of 1 input, 2 hidden units, 1 output, with Gaussian noise and randomly sampled for 1000 trials

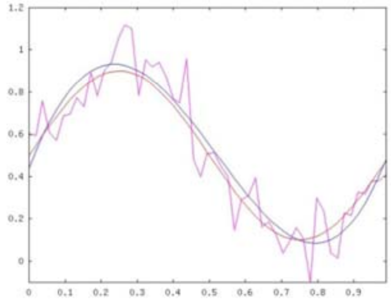


Figure 3 – Noisy Sine Wave

In Figure 3 the red sine wave is the true denoised signal, the noisy red line is the actual input, and the blue line is the prediction of the network.

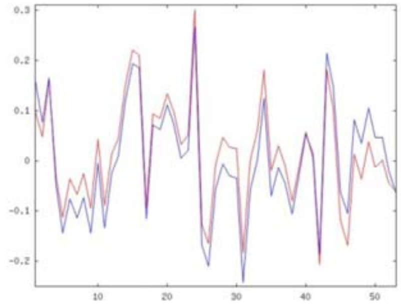


Figure 4 – Noise only (Sine Wave)

In Figure 4 the red curve is the true noise in the signal, and the blue line is the actual input less the prediction of the network, i.e, is the modeled noise resulting from the network prediction.

Decaying Exponential

This experiment is for a noisy decaying exponential curve and was pattern matched with a network consisting of 1 input, 1 hidden unit, 1 output, with Gaussian noise.

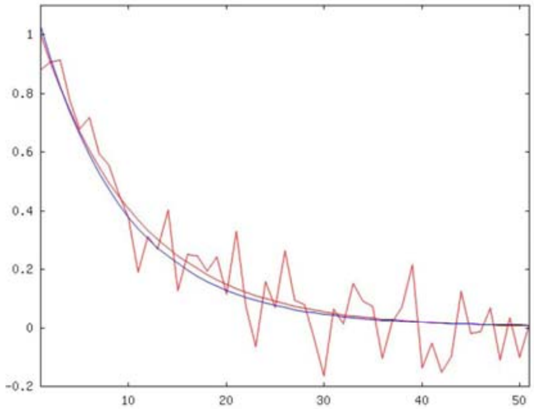


Figure 5 – Noisy Decaying Exponential

In Figure 5 the red line is the true denoised signal, the noisy red line is the actual input, and the blue line is the prediction of the network.

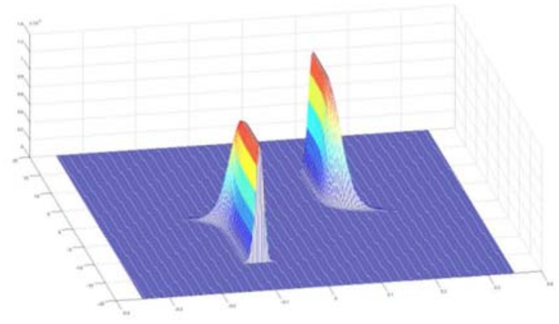


Figure 6 – Posterior for the Two Weights for Decaying Exponential

In Figure 6 we have the posterior distribution for the two hidden weights for the decaying exponential problem determined by grid sampling. There are clearly two branches of significant probability above some floor.

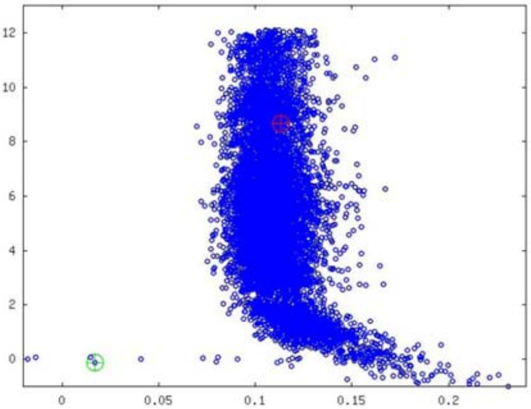


Figure 7 – Metropolis Markov Chain Samples for Two Weights for Decaying Exponential

In Figure 7 we have the sample points from the Metropolis algorithm for the two hidden weights for the decaying exponential problem. Note the correspondence between Figures 6 and 7. The Metropolis algorithm has found one branch of the solution depicted in Figure 6.

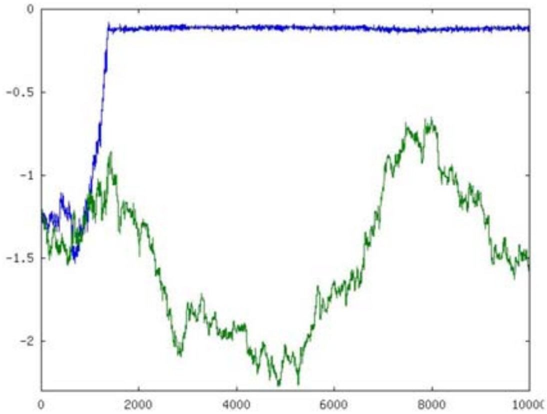


Figure 8 – Metropolis Markov Chain Process for Two Weights in Decaying Exponential

In Figure 8 we show the processes for the sample for the two weights of the decaying exponential fit. Note that the process for one of the parameters has found the correct value after a warm up of approximately 1500 steps in the Markov Chain. The other parameters' process is more of a random walk due to its wide range of acceptable values (compare with Figures 6 and 7).

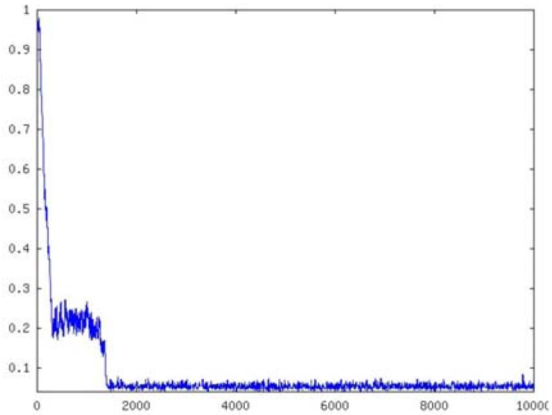


Figure 9 – Metropolis Markov Chain Process for Noise Parameter

In Figure 9 we show the process for the sample for the noise component parameter of the decaying exponential fit. We note that the process for the noise parameter has found the correct value of $\mathcal{N}(0,0.05^2)$ after a warm up of approximately 1500 steps in the Markov Chain.

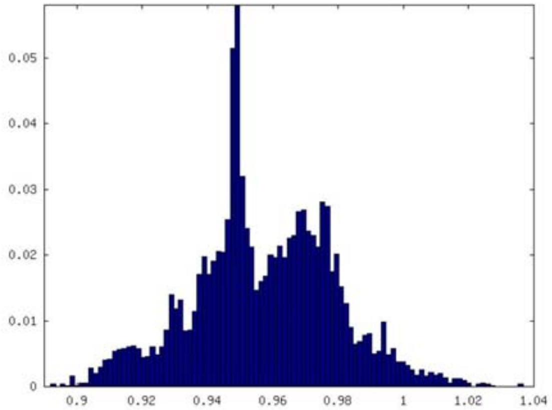


Figure 10 – Uncertainty in Network Prediction

In Figure 10 we show the normalized probability distribution from the resultant Markov Chain for the first output point. The denoised output for the first point is actually 1.

Concrete Problem

This problem known as the Concrete Problem is from the Machine Learning Database at UC Irvine and consists of 7 inputs and 3 outputs. It was processed with a network of 32 hidden units and only 500 MCMC samples after a warm up of 500 samples. Also included in the sampling for this problem were all the components of the full covariance matrix for the outputs including the off-diagonal components (ref. equations 9,10)

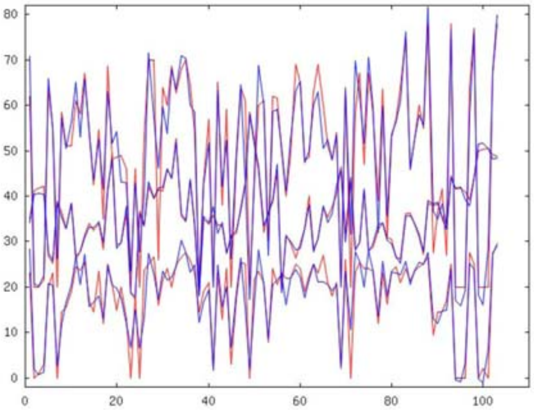


Figure 11 – Comparison between network output (blue) and training input (red) for 3 outputs in Concrete problem

In Figure 11 we compare the training data (red) with the network prediction (blue) for the concrete problem

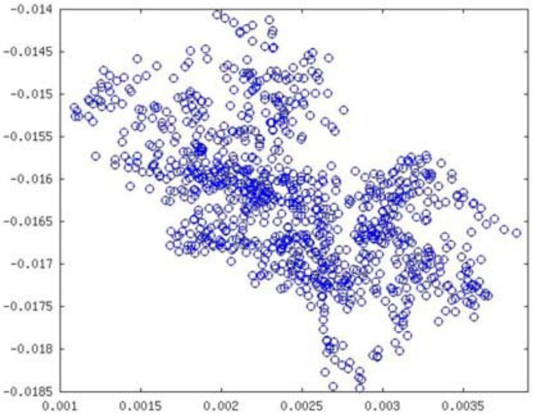


Figure 12 – Typical Metropolis Samples for Two of the Weights in Concrete problem

4.0 DISCUSSION / CONCLUSIONS

The selected experiments showed good responses of the methodology for surprisingly few numbers of trials. In cases where the actual strength of the noise component of the signal was known, the method reliably inferred a value very close to the actual value, with small variance on repeated trials. In all experiments we note the fairly rapid convergence of the chain to promising potential solutions starting from randomize initial locations. The implemented MCMC sampling algorithm was admittedly crude and typically achieved acceptance fractions of between 10% to 15% well below that recommended by the literature (25% to 50%).

5.0 REFERENCES

[1] G. Cybenko. Approximations by superpositions of sigmoidal functions. Mathematics of Control, Signals, and Systems, 2: no. 4 pp. 303-314. 1989.

[2] K. Hornik: Approximation Capabilities of Multilayer Feedforward Networks. Neural Networks, vol. 4, 1991.

[3] R. T. Cox, The Algebra of Probable Inference. The Johns Hopkins Press, 1961

[4] E. T. Jaynes, Probability Theory: The Logic of Science, Cambridge University Press., 2003

[5] Jeffreys, H An Invariant Form for the Prior Probability in Estimation Problems Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, Vol. 186, No. 1007 (Sep. 24, 1946), pp. 453-461

[6] Jeffreys, H.. Theory of Probability, third edition, Oxford University Press, 1961

[7] Roberts, G.O.; Gelman, A.; Gilks, W.R. (1997). "Weak convergence and optimal scaling of random walk Metropolis algorithms", *Ann. Appl. Probab.* 7 (1): 110–120

[8] Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. (1953). "Equations of State Calculations by Fast Computing Machines", *Journal of Chemical Physics* 21 (6): 1087–1092

[9] Hastings, W.K. (1970). "Monte Carlo Sampling Methods Using Markov Chains and Their Applications", *Biometrika* **57** (1): 97–109

[10] Sivia, D.S. "Data Analysis: A Bayesian Tutorial". *Oxford University Press,* 1996, ch 6.

[11] Gregory, P. "Bayesian Logical Data Analysis for the Physical Sciences", *Cambidge University Press,* 2005, ch 3.

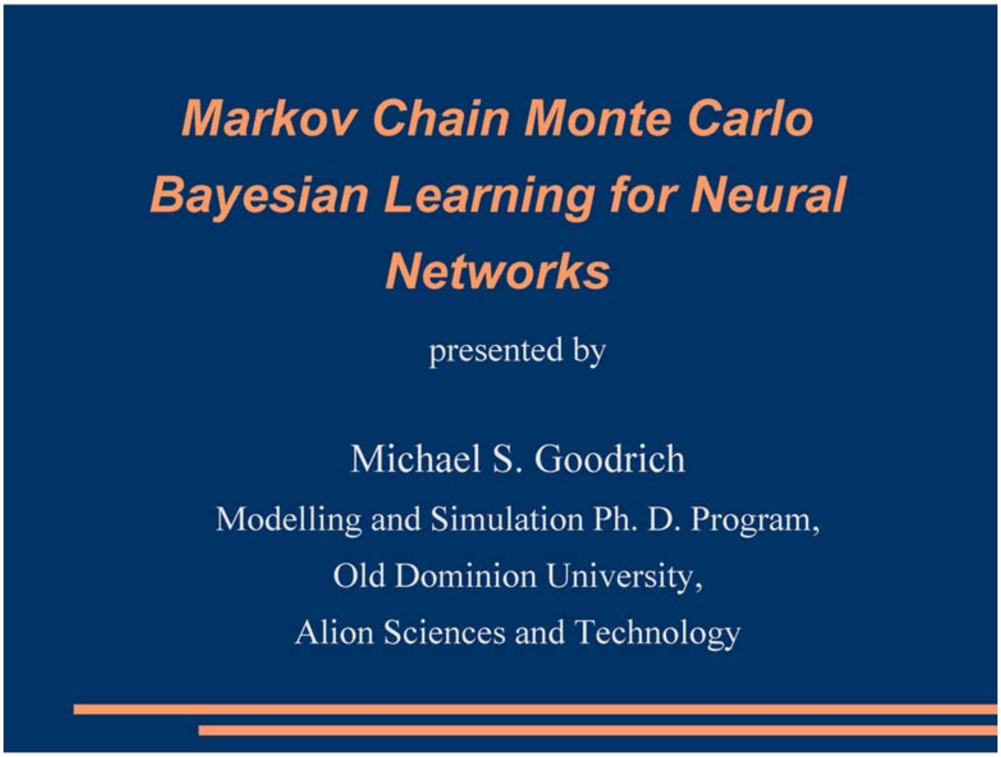
[12] Bishop, C.M. "Pattern Recognition and Machine Learning", *Springer Press,* 2006, p. 10

[13] Lee, P. "Bayesian Statistics: An Intorduction", *Oxford University Press,* 1989, pp 90-97.

[14] Gamerman, D., Lopes, H., "Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference", 2^{nd} Ed., , *Chapman & Hall/CRC,* 2006.

6.0 ACKNOWLEDGMENT(S)

We would like to thank Dr. Jiang Li, Old Dominion University, for being supportive of this line of research, and for suggesting that a paper along these lines be presented at this conference.



Markov Chain Monte Carlo Bayesian Learning for Neural Networks

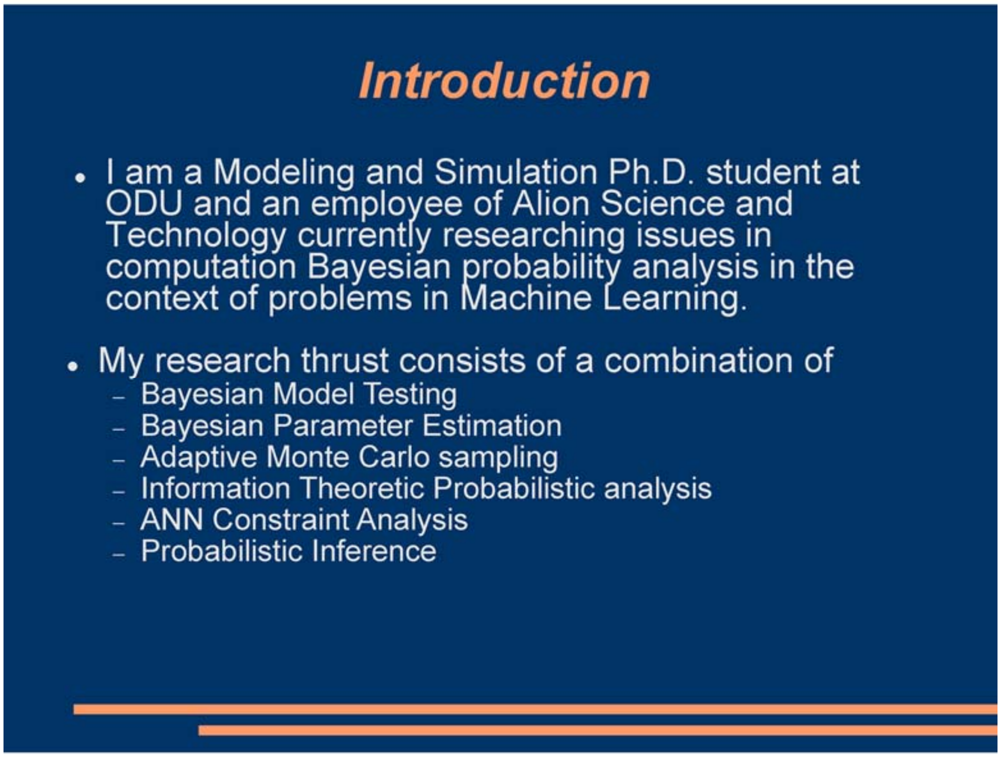
presented by

Michael S. Goodrich

Modelling and Simulation Ph. D. Program,

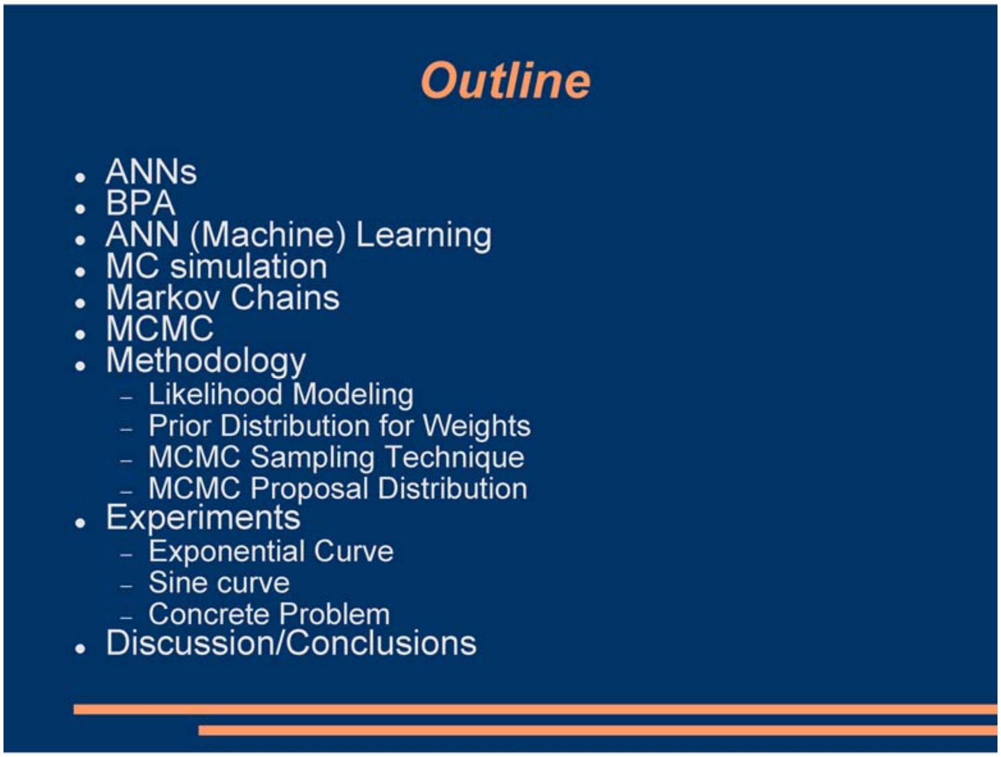
Old Dominion University,

Alion Sciences and Technology



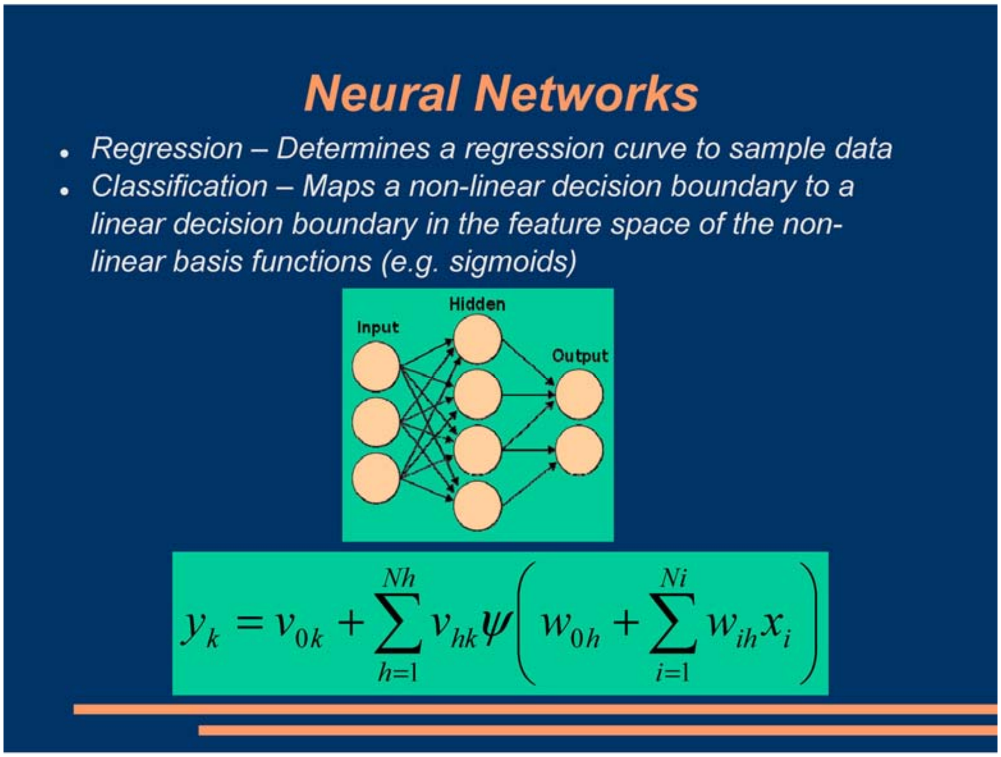
Introduction

- I am a Modeling and Simulation Ph.D. student at ODU and an employee of Alion Science and Technology currently researching issues in computation Bayesian probability analysis in the context of problems in Machine Learning.
- My research thrust consists of a combination of
  - Bayesian Model Testing
  - Bayesian Parameter Estimation
  - Adaptive Monte Carlo sampling
  - Information Theoretic Probabilistic analysis
  - ANN Constraint Analysis
  - Probabilistic Inference



Outline

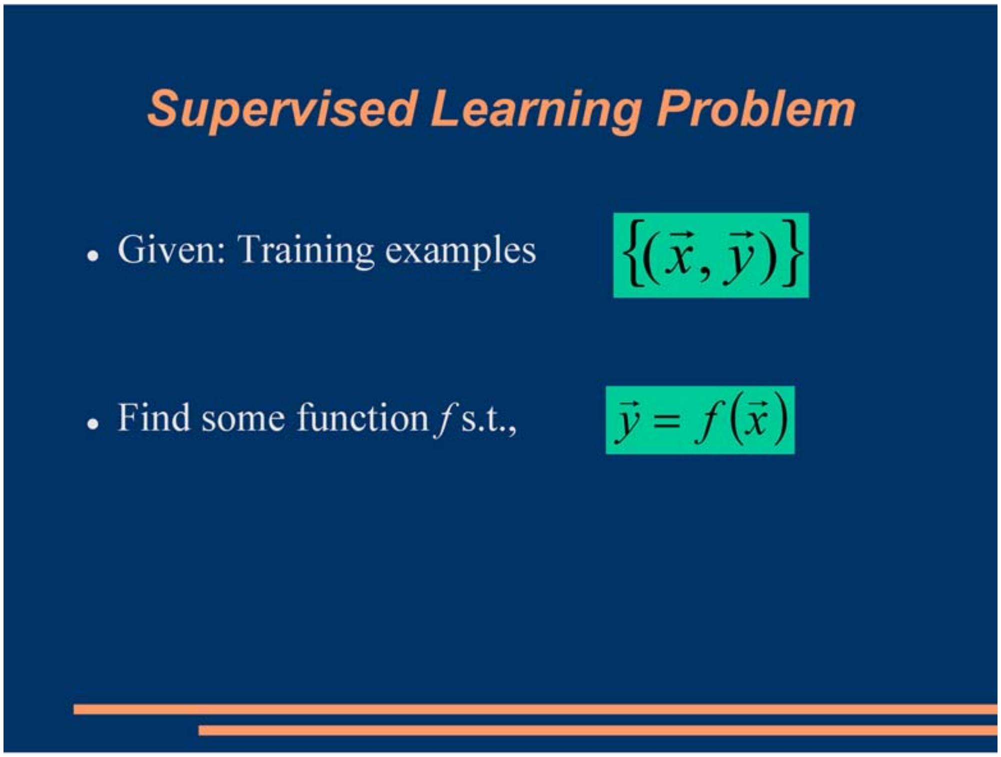
- ANNs
- BPA
- ANN (Machine) Learning
- MC simulation
- Markov Chains
- MCMC
- Methodology
  - Likelihood Modeling
  - Prior Distribution for Weights
  - MCMC Sampling Technique
  - MCMC Proposal Distribution
- Experiments
  - Exponential Curve
  - Sine curve
  - Concrete Problem
- Discussion/Conclusions



Neural Networks

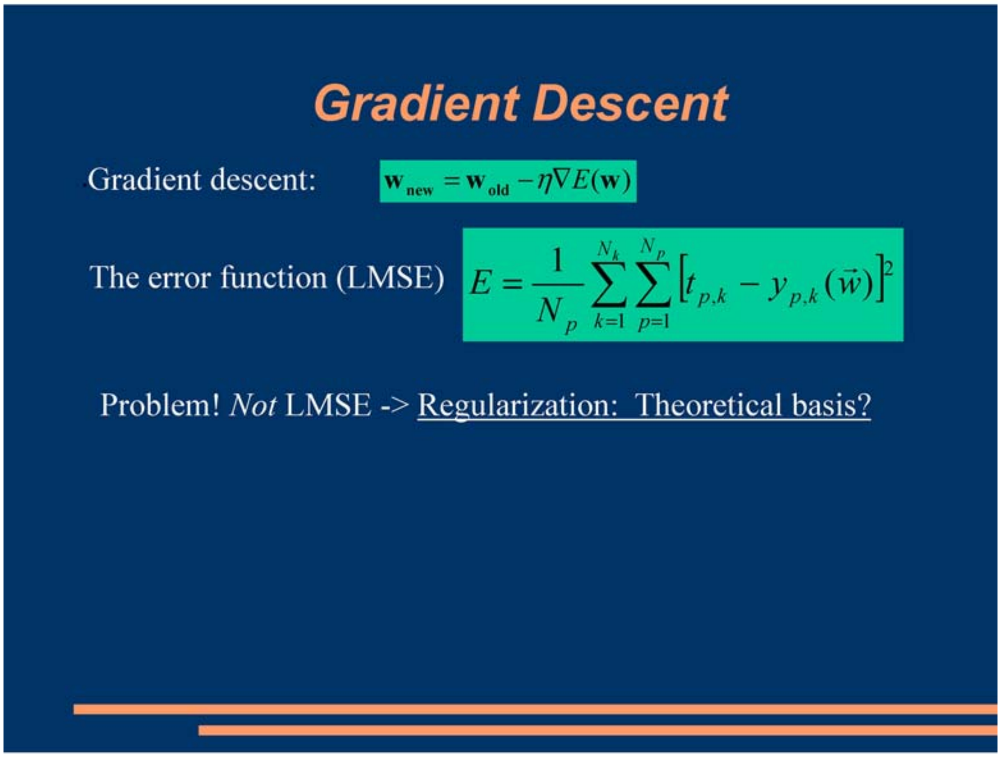
- *Regression – Determines a regression curve to sample data*
- *Classification – Maps a non-linear decision boundary to a linear decision boundary in the feature space of the non-linear basis functions (e.g. sigmoids)*

$$y_k = v_{0k} + \sum_{h=1}^{Nh} v_{hk}\psi\left(w_{0h} + \sum_{i=1}^{Ni} w_{ih}x_i \right)$$



Supervised Learning Problem

- Given: Training examples $\{(\vec{x}, \vec{y})\}$
- Find some function *f* s.t., $\vec{y} = f(\vec{x})$

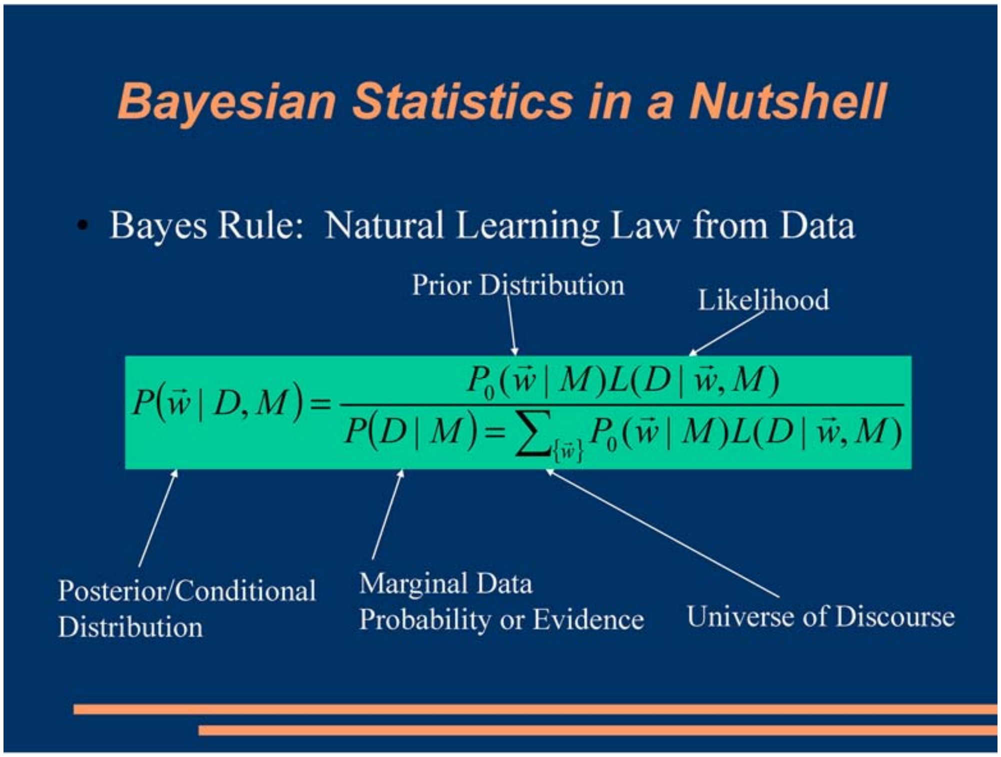


Gradient Descent

Gradient descent: $\mathbf{w}_{\text{new}} = \mathbf{w}_{\text{old}} - \eta \nabla E(\mathbf{w})$

The error function (LMSE) $$E = \frac{1}{N_p} \sum_{k=1}^{N_k} \sum_{p=1}^{N_p} \left[t_{p,k} - y_{p,k}(\vec{w}) \right]^2$$

Problem! *Not* LMSE -> Regularization: Theoretical basis?



Bayesian Statistics in a Nutshell

- Bayes Rule: Natural Learning Law from Data

Prior Distribution

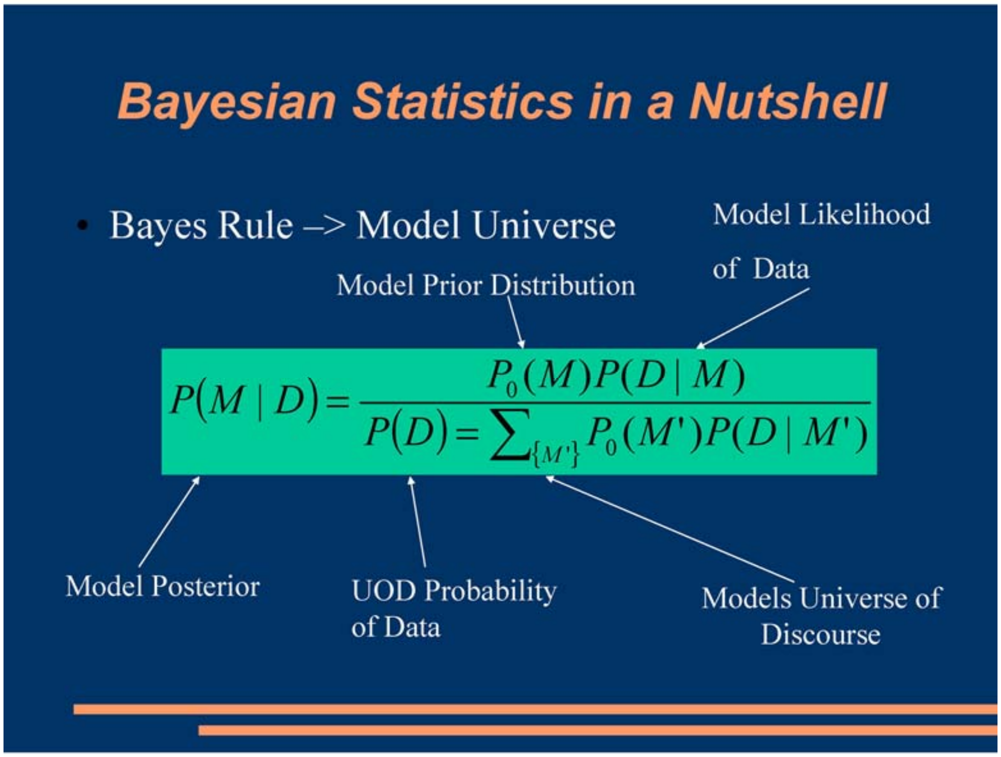
Likelihood

$$P(\vec{w} \mid D, M) = \frac{P_0(\vec{w} \mid M) L(D \mid \vec{w}, M)}{P(D \mid M) = \sum_{\{\vec{w}\}} P_0(\vec{w} \mid M) L(D \mid \vec{w}, M)}$$

Posterior/Conditional Distribution

Marginal Data Probability or Evidence

Universe of Discourse



Bayesian Statistics in a Nutshell

- Bayes Rule –> Model Universe

Model Likelihood of Data

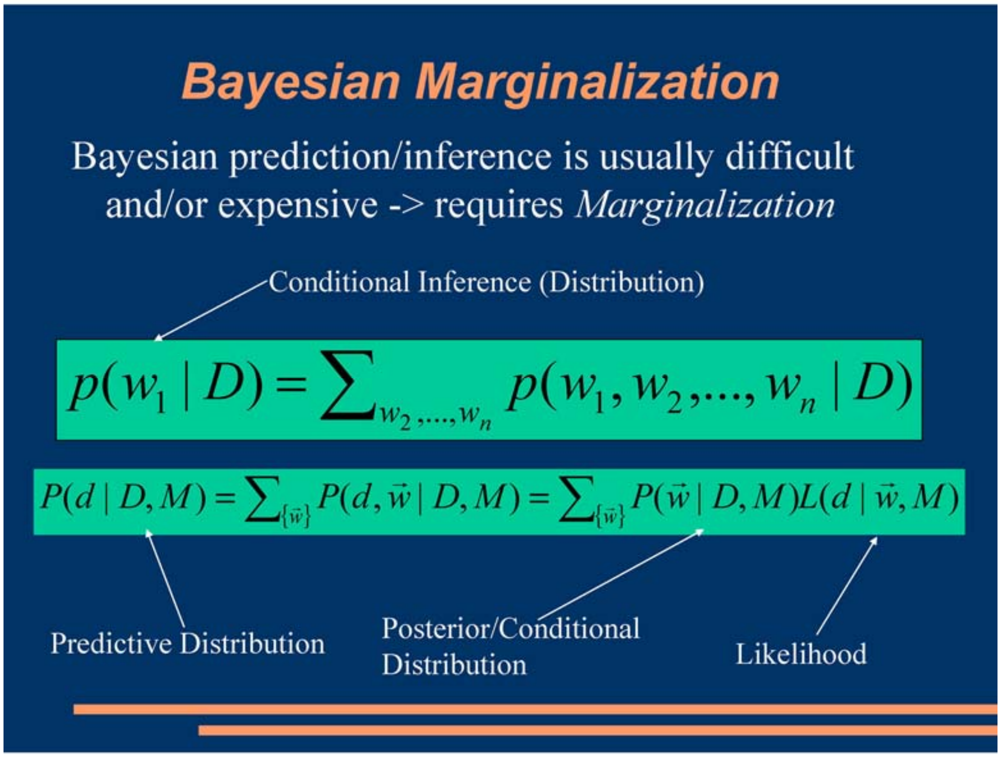
Model Prior Distribution

$$P(M \mid D) = \frac{P_0(M) P(D \mid M)}{P(D) = \sum_{\{M'\}} P_0(M') P(D \mid M')}$$

Model Posterior

UOD Probability of Data

Models Universe of Discourse



Bayesian Marginalization

Bayesian prediction/inference is usually difficult and/or expensive -> requires *Marginalization*

Conditional Inference (Distribution)

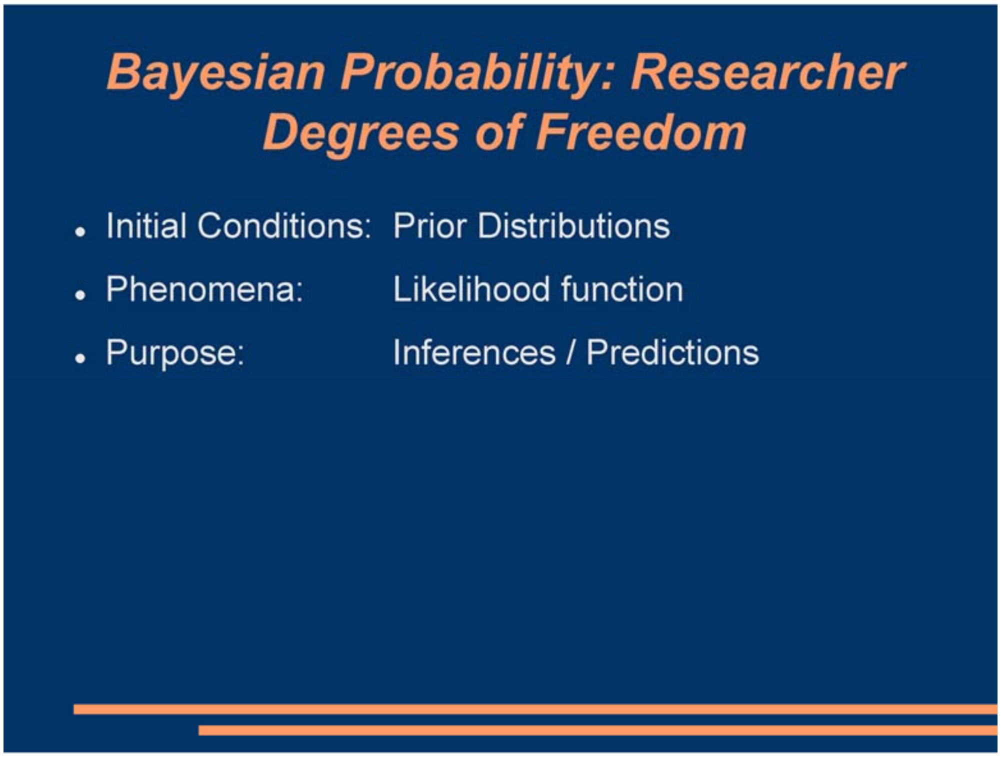
$$p(w_1 \mid D) = \sum_{w_2,\ldots,w_n} p(w_1, w_2,\ldots, w_n \mid D)$$

$$P(d \mid D, M) = \sum_{\{\vec{w}\}} P(d, \vec{w} \mid D, M) = \sum_{\{\vec{w}\}} P(\vec{w} \mid D, M) L(d \mid \vec{w}, M)$$

Predictive Distribution

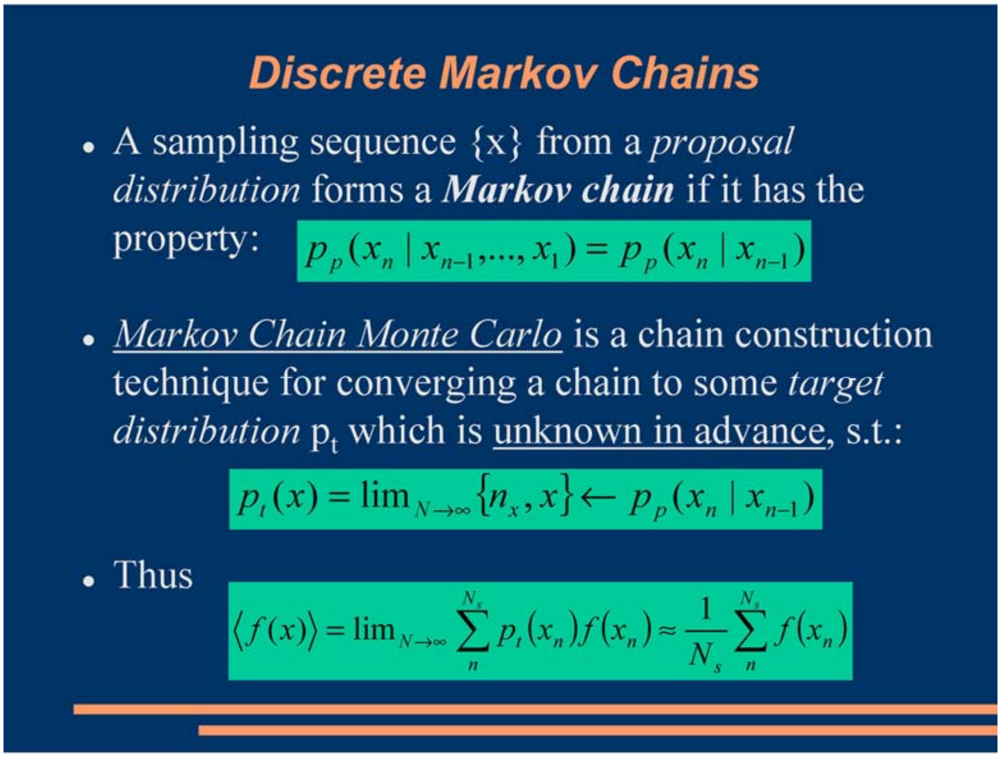
Posterior/Conditional Distribution

Likelihood



Bayesian Probability: Researcher Degrees of Freedom

Initial Conditions:	Prior Distributions
Phenomena:	Likelihood function
Purpose:	Inferences / Predictions



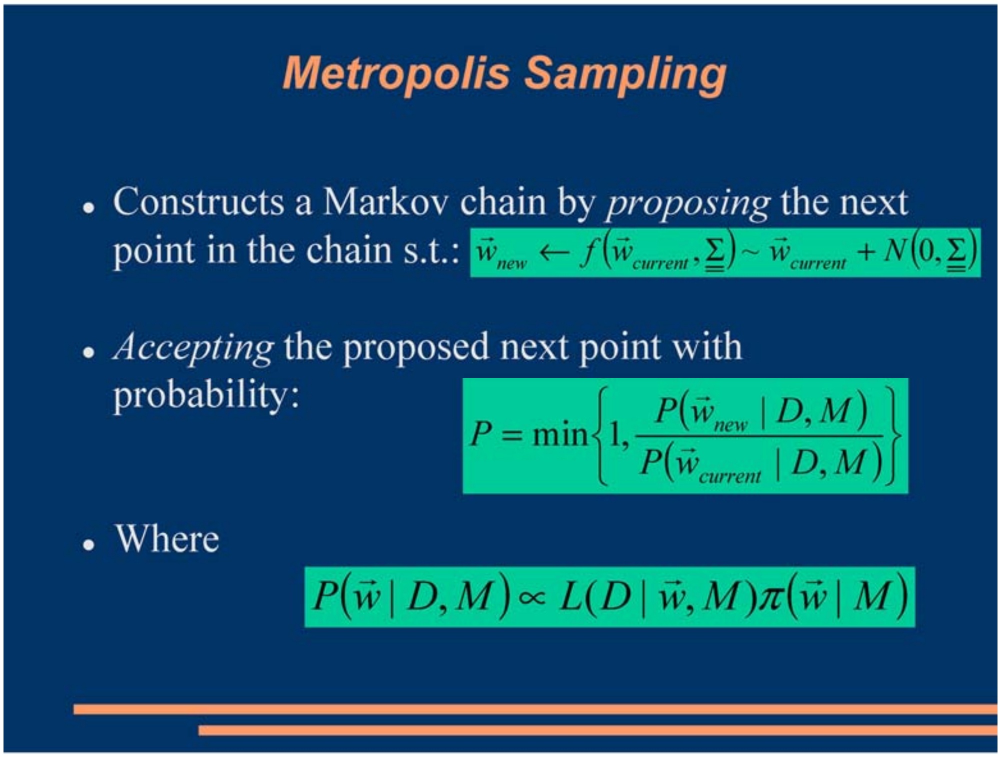
Discrete Markov Chains

- A sampling sequence {x} from a *proposal distribution* forms a ***Markov chain*** if it has the property: $p_p(x_n \mid x_{n-1},\ldots,x_1) = p_p(x_n \mid x_{n-1})$
- *Markov Chain Monte Carlo* is a chain construction technique for converging a chain to some *target distribution* p_t which is unknown in advance, s.t.:

$$p_t(x) = \lim_{N\to\infty}\{n_x, x\} \leftarrow p_p(x_n \mid x_{n-1})$$

- Thus

$$\langle f(x)\rangle = \lim_{N\to\infty}\sum_{n}^{N_s} p_t(x_n) f(x_n) \approx \frac{1}{N_s}\sum_{n}^{N_s} f(x_n)$$



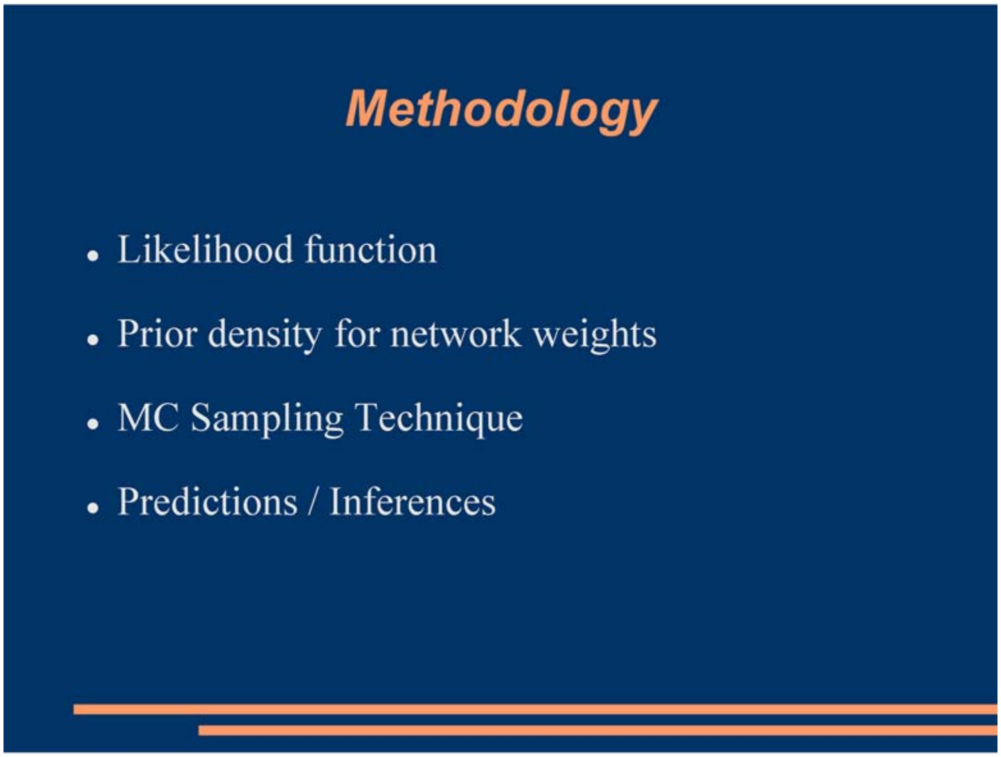
Metropolis Sampling

- Constructs a Markov chain by *proposing* the next point in the chain s.t.: $\vec{w}_{new} \leftarrow f\left(\vec{w}_{current}, \underline{\underline{\Sigma}}\right) \sim \vec{w}_{current} + N\left(0, \underline{\underline{\Sigma}}\right)$
- *Accepting* the proposed next point with probability:

$$P = \min\left\{1, \frac{P\left(\vec{w}_{new} \mid D, M\right)}{P\left(\vec{w}_{current} \mid D, M\right)}\right\}$$

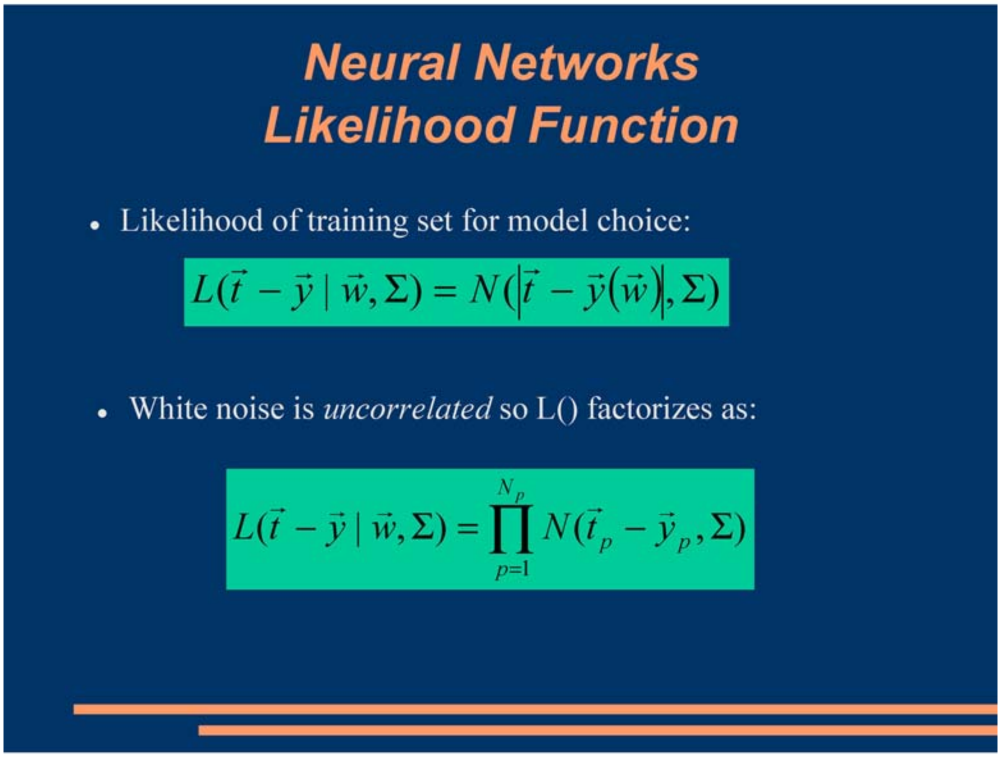
- Where

$$P\left(\vec{w} \mid D, M\right) \propto L(D \mid \vec{w}, M)\pi\left(\vec{w} \mid M\right)$$



Methodology

- Likelihood function
- Prior density for network weights
- MC Sampling Technique
- Predictions / Inferences



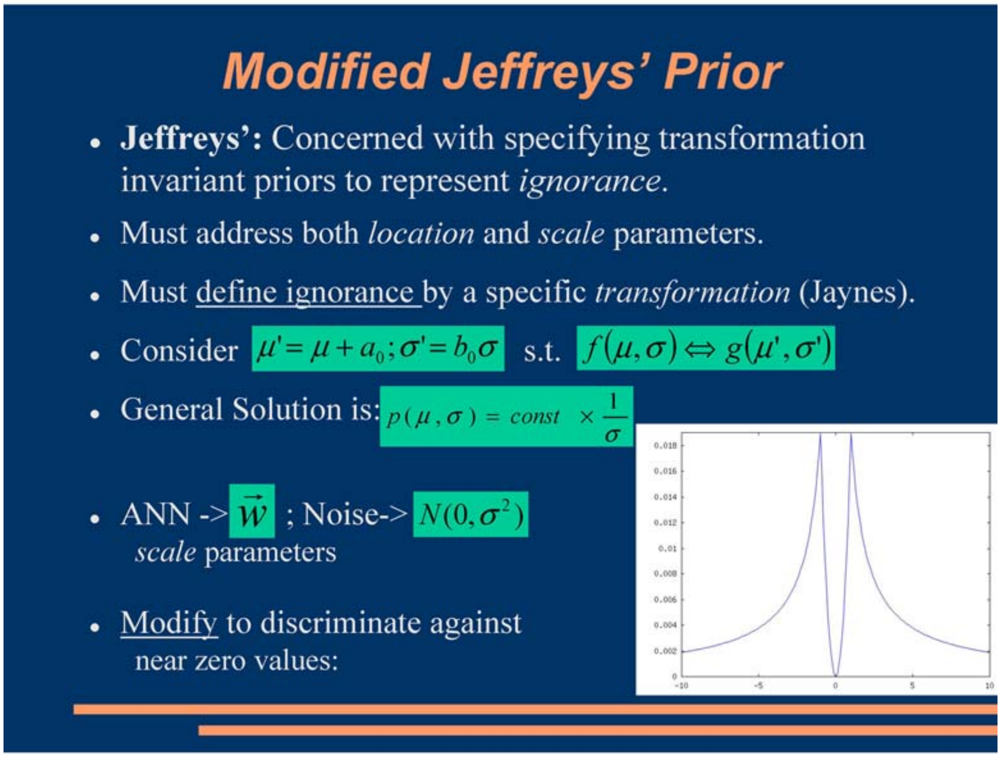
Neural Networks Likelihood Function

- Likelihood of training set for model choice:

$$L(\vec{t}-\vec{y}\mid\vec{w},\Sigma)=N(\left|\vec{t}-\vec{y}(\vec{w})\right|,\Sigma)$$

- White noise is *uncorrelated* so L() factorizes as:

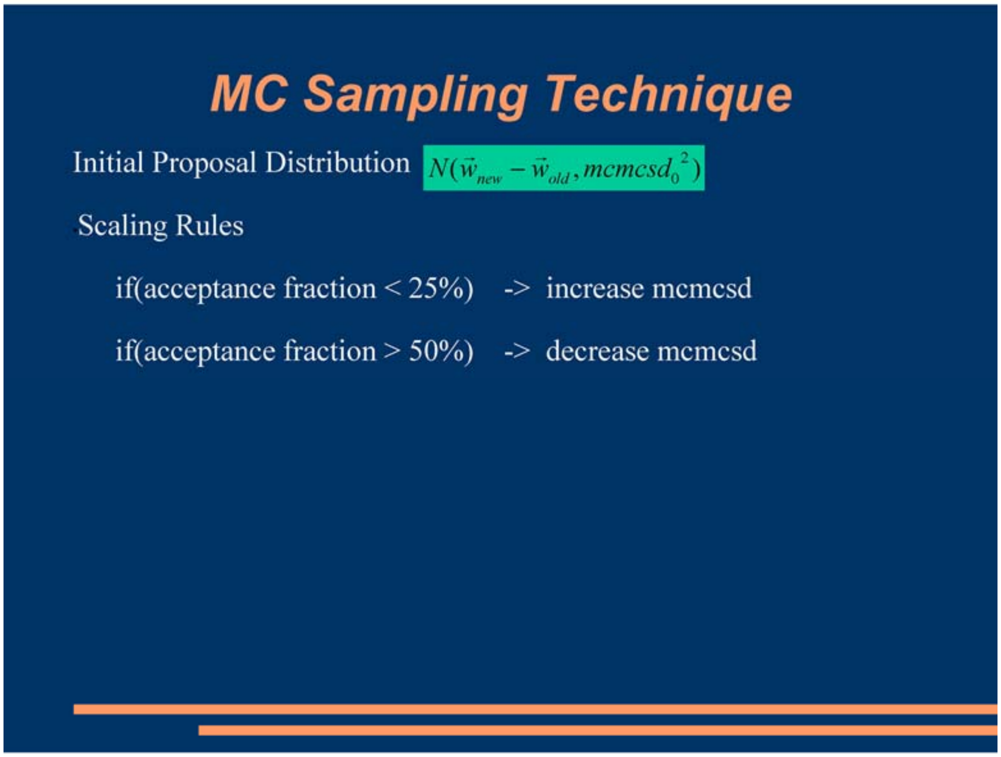
$$L(\vec{t}-\vec{y}\mid\vec{w},\Sigma)=\prod_{p=1}^{N_p}N(\vec{t}_p-\vec{y}_p,\Sigma)$$



Modified Jeffreys' Prior

- **Jeffreys':** Concerned with specifying transformation invariant priors to represent *ignorance*.
- Must address both *location* and *scale* parameters.
- Must define ignorance by a specific *transformation* (Jaynes).
- Consider $\mu' = \mu + a_0 ; \sigma' = b_0 \sigma$ s.t. $f(\mu, \sigma) \Leftrightarrow g(\mu', \sigma')$
- General Solution is: $p(\mu, \sigma) = const \times \frac{1}{\sigma}$
- ANN -> $\vec{w}$; Noise-> $N(0, \sigma^2)$
  scale parameters
- Modify to discriminate against
  near zero values:

0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
-10
-5
0
5
10



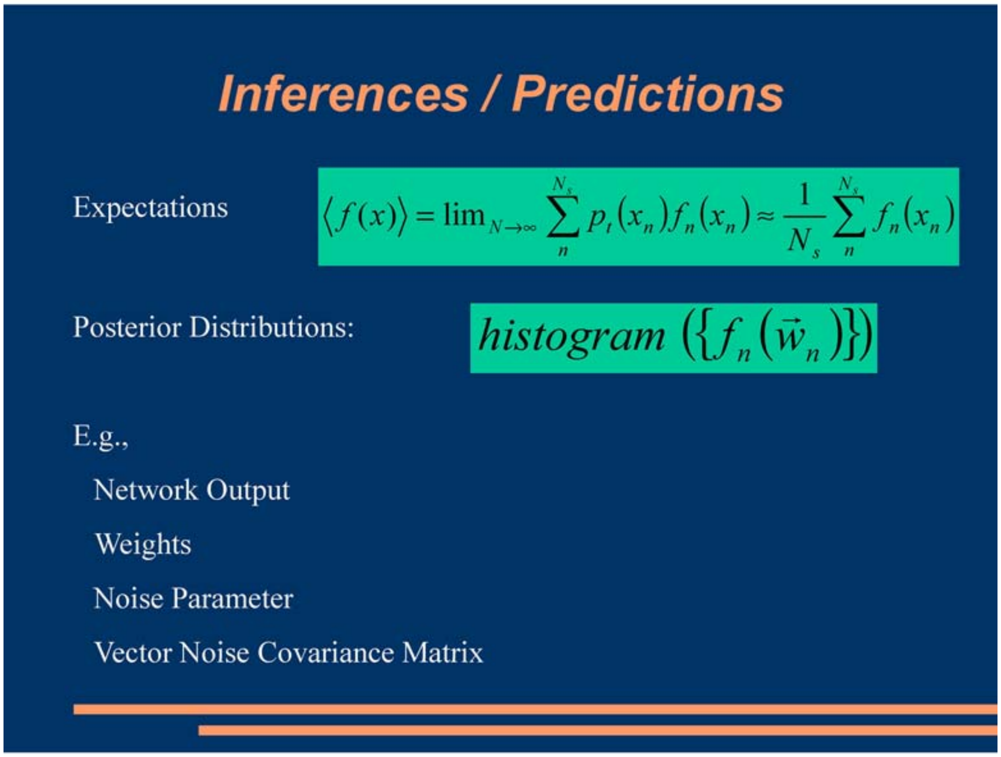
MC Sampling Technique

Initial Proposal Distribution $N(\vec{w}_{new} - \vec{w}_{old}, mcmcsd_0^{\;2})$

Scaling Rules

if(acceptance fraction < 25%) -> increase mcmcsd

if(acceptance fraction > 50%) -> decrease mcmcsd



Inferences / Predictions

Expectations $\langle f(x)\rangle = \lim_{N\to\infty} \sum_{n}^{N_s} p_t(x_n) f_n(x_n) \approx \frac{1}{N_s}\sum_{n}^{N_s} f_n(x_n)$

Posterior Distributions: $histogram\left(\left\{f_n(\vec{w}_n)\right\}\right)$

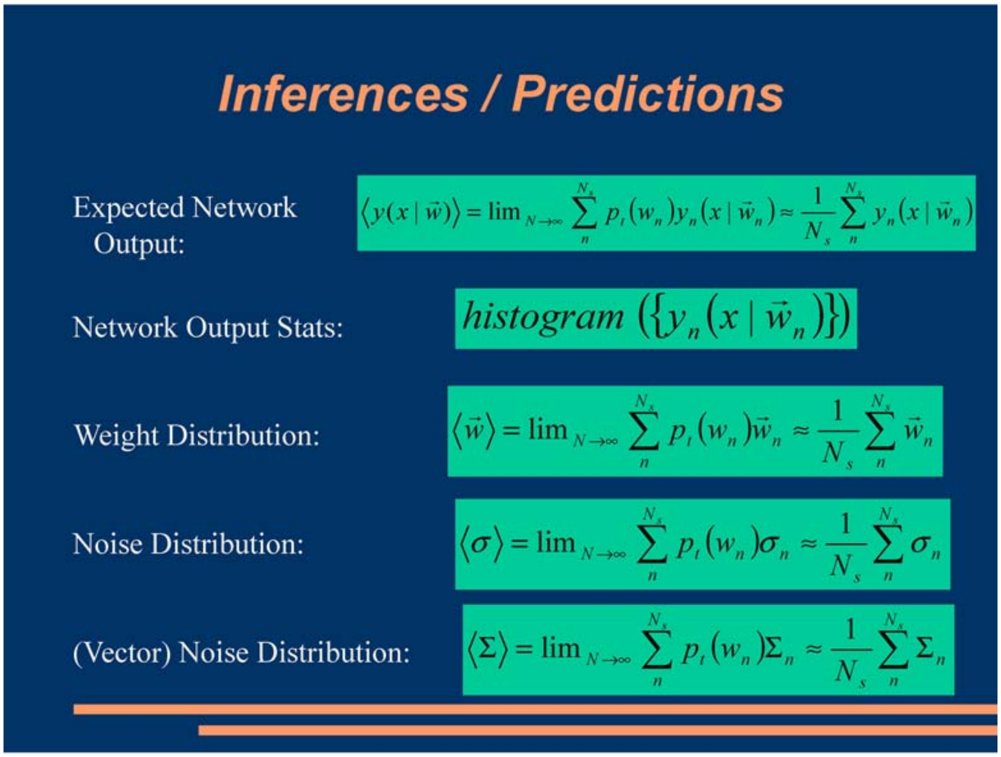
E.g.,

Network Output

Weights

Noise Parameter

Vector Noise Covariance Matrix



Inferences / Predictions

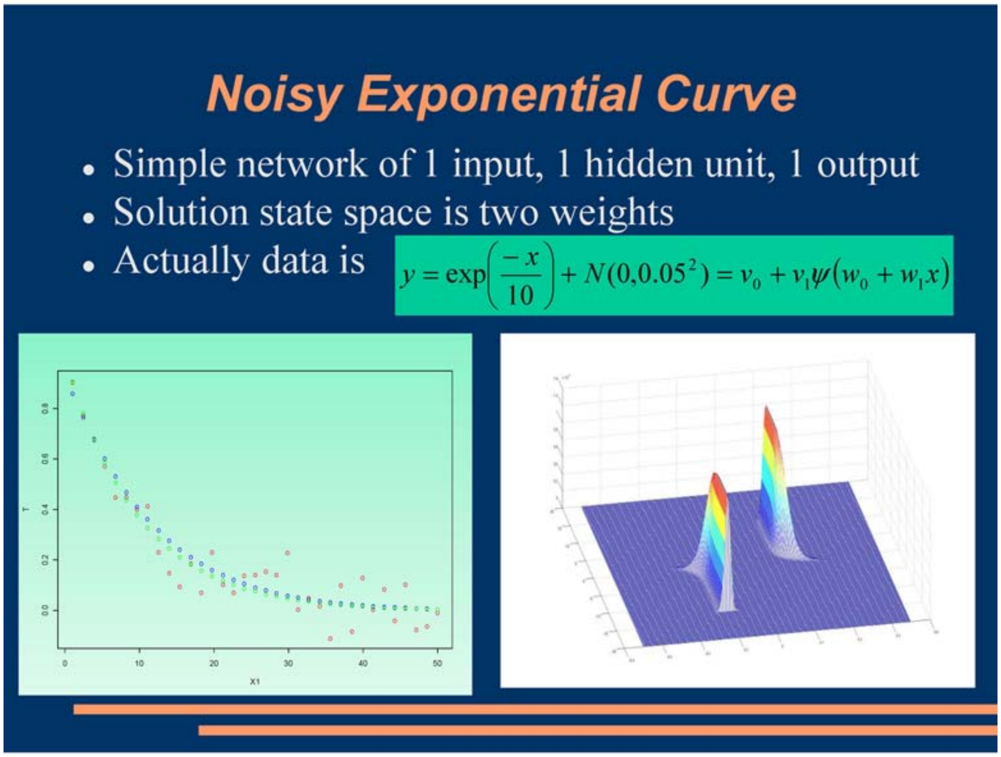
Expected Network Output: $\langle y(x\mid\vec{w})\rangle = \lim_{N\to\infty} \sum_{n}^{N_s} p_t(w_n) y_n(x\mid\vec{w}_n) \approx \frac{1}{N_s}\sum_{n}^{N_s} y_n(x\mid\vec{w}_n)$

Network Output Stats: $histogram\left(\left\{y_n(x\mid\vec{w}_n)\right\}\right)$

Weight Distribution: $\langle\vec{w}\rangle = \lim_{N\to\infty} \sum_{n}^{N_s} p_t(w_n)\vec{w}_n \approx \frac{1}{N_s}\sum_{n}^{N_s}\vec{w}_n$

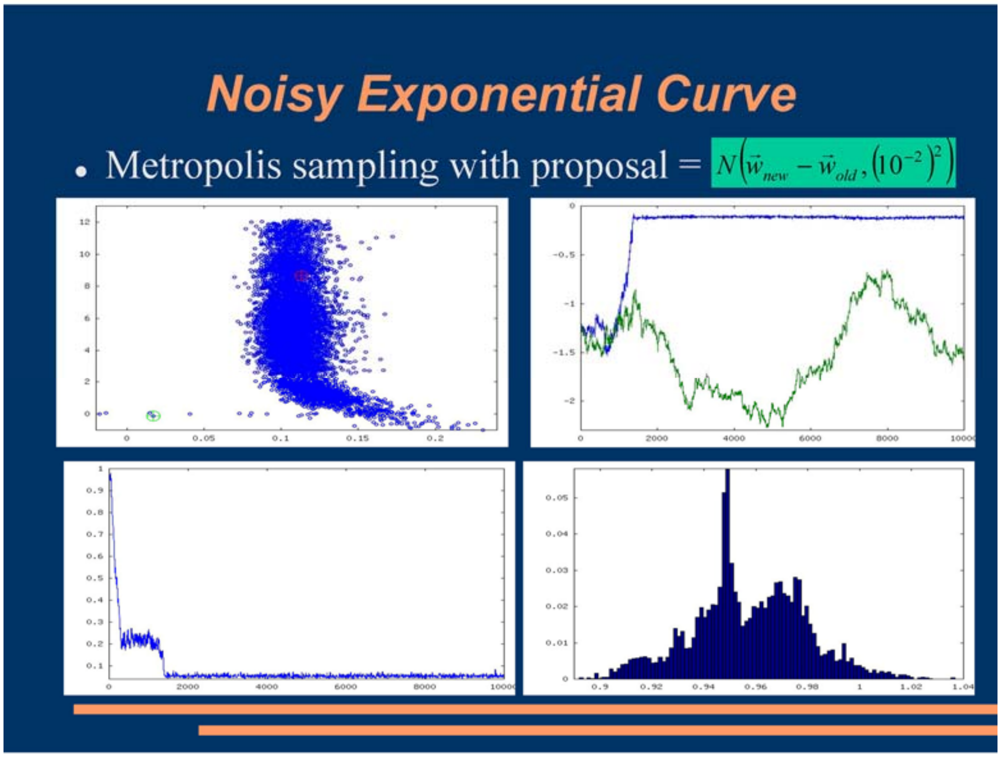
Noise Distribution: $\langle\sigma\rangle = \lim_{N\to\infty} \sum_{n}^{N_s} p_t(w_n)\sigma_n \approx \frac{1}{N_s}\sum_{n}^{N_s}\sigma_n$

(Vector) Noise Distribution: $\langle\Sigma\rangle = \lim_{N\to\infty} \sum_{n}^{N_s} p_t(w_n)\Sigma_n \approx \frac{1}{N_s}\sum_{n}^{N_s}\Sigma_n$



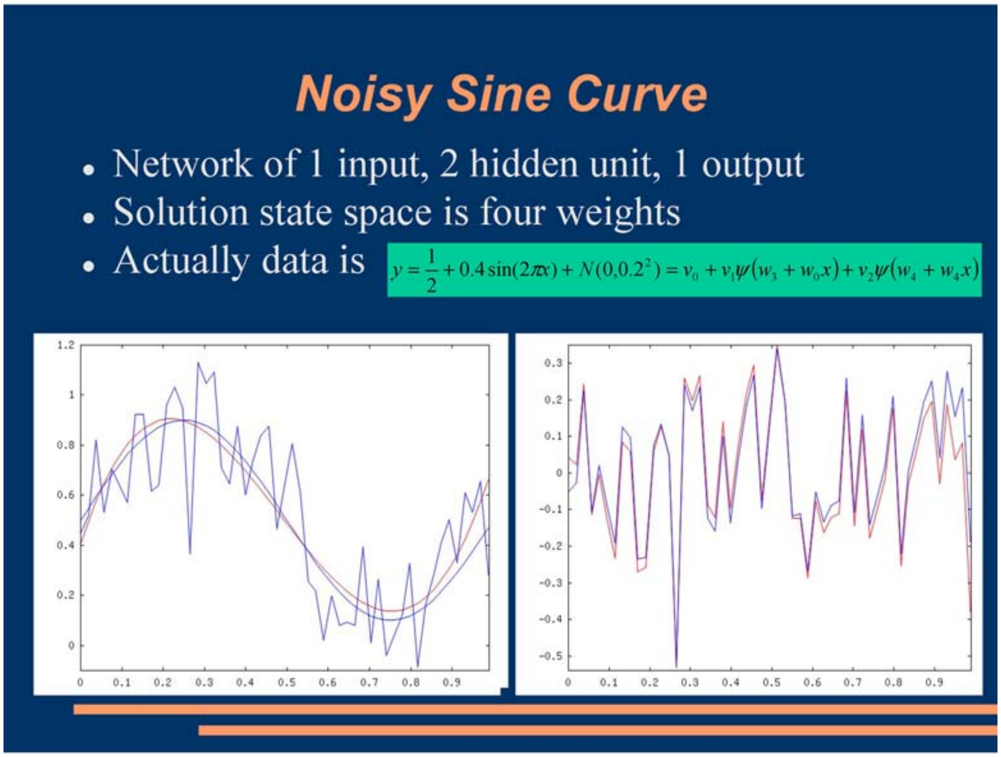
Noisy Exponential Curve

- Simple network of 1 input, 1 hidden unit, 1 output
- Solution state space is two weights
- Actually data is $y = \exp\left(\frac{-x}{10}\right) + N(0,0.05^2) = v_0 + v_1\psi\left(w_0 + w_1 x\right)$



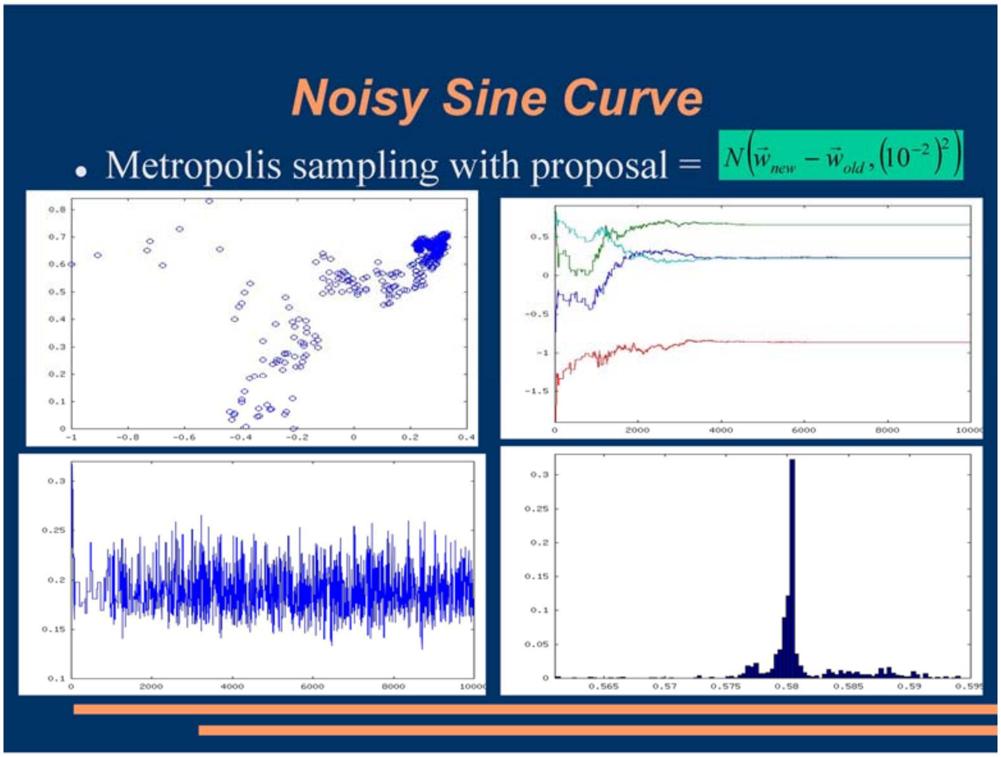
Noisy Exponential Curve

- Metropolis sampling with proposal = $N\left(\vec{w}_{new} - \vec{w}_{old}, \left(10^{-2}\right)^2\right)$



Noisy Sine Curve

- Network of 1 input, 2 hidden unit, 1 output
- Solution state space is four weights
- Actually data is $y = \frac{1}{2} + 0.4\sin(2\pi x) + N(0,0.2^2) = v_0 + v_1\psi(w_3 + w_0 x) + v_2\psi(w_4 + w_4 x)$

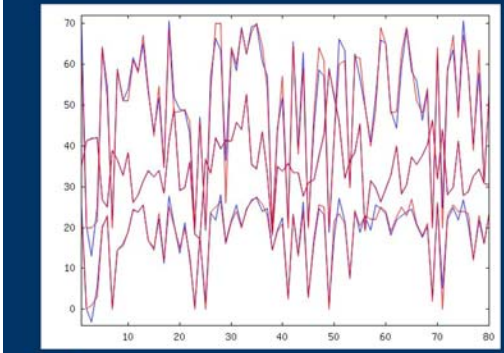


Noisy Sine Curve

- Metropolis sampling with proposal = $N\left(\vec{w}_{new} - \vec{w}_{old}, \left(10^{-2}\right)^2\right)$

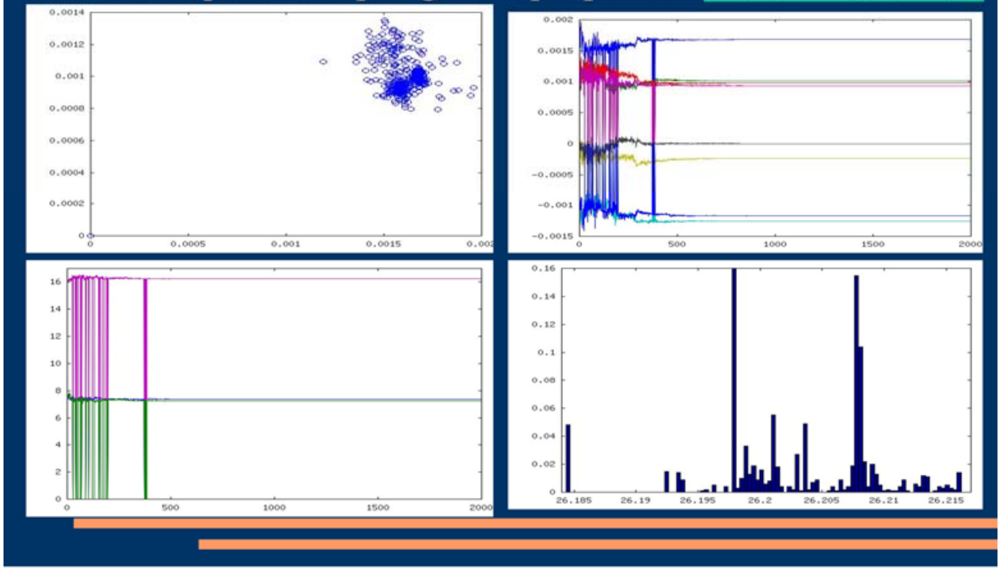
Concrete Problem

- Network of 7 inputs, 64 hidden units, 3 outputs
- Solution state space is (7+1)64 = 512 weights
- Actually data is real-world -> “noise” component unknown



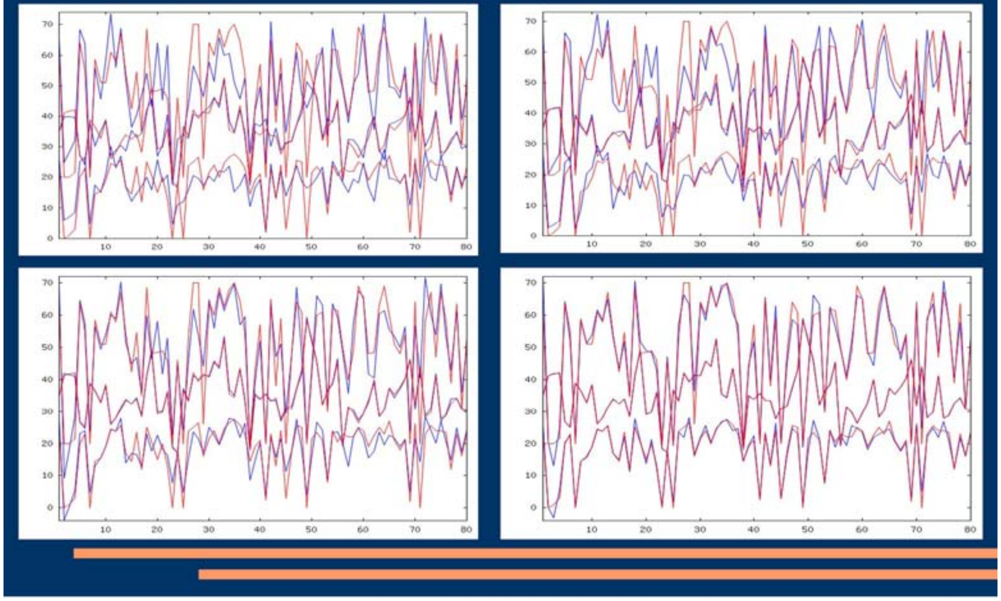
Concrete Problem

- Metropolis sampling with proposal = $N\left(\vec{w}_{new} - \vec{w}_{old}, \left(10^{-3}\right)^2\right)$



Concrete Problem

- Comparative Predictions of 8,16,32,64 hidden units



Discussion / Conclusions

- A promising start!
- Surprisingly good results for small sample sizes
- Other options for priors:
  - Constraints
  - Non Linear solvers
- Proposal Density Scaling Issues
- Regularization: Model Selection appears best.

2.17 Enabling Rapid Naval Architecture Design Space Exploration

Enabling Rapid Naval Architecture Design Space Exploration

Michael A. Mueller, Stephane Dufresne, Santiago Balestrini-Robinson, and Dimitri Mavris
Georgia Institute of Technology
mmueller@asdl.gatech.edu stephane.dufresne@asdl.gatech.edu
santiago.balestrini@asdl.gatech.edu dmavris@asdl.gatech.edu

Well accepted conceptual ship design tools can be used to explore a design space, but more precise results can be found using detailed models in full-feature computer aided design programs. However, defining a detailed model can be a time intensive task, and hence there is an incentive for time sensitive projects to use conceptual design tools to explore the design space. In this project, the combination of advanced aerospace systems design methods and an accepted conceptual design tool facilitates the creation of a tool that enables the user to not only visualize ship geometry but also determine design feasibility and estimate the performance of a design.

1.0 INTRODUCTION

The practice of naval architecture parallels the field of aerospace engineering in many ways; both involve a calculated balance of resistance, power, and weight, and both require an early recognition of the desired capabilities in the design of the respective craft. It is in this latter parallel where systems engineering methods can be very useful throughout the design process. An important part of the modern systems engineering process involves modeling the craft to better understand not only the interactions between the internal subsystems but also the interactions between the craft and the environment. In both aerospace engineering and naval architecture, physics-based models, models defined by realistic physics and processes, can be used as models to help the designers develop balanced and effective solutions to the problems posed by the design requirements.

Physics-based models allow for the prediction of performance and can be very precise. Work by Jiang, Forstell, Lavis, and Ritter demonstrated the power of the ship design program PASS, Parametric Analysis of Ship Systems, a physics-based design modeling software, to accurately predict the design parameters of the CG-47 cruiser and carry out additional performance optimization analysis [1]. Though physics-based models can be useful tool in modern design methodology, the ability to design and analyze craft can be further advanced through the use of three-dimensional, 3-D, product modeling software. Such software can allow for the creation of electronic mock-ups, reduce the number of design and rework errors, and allow for concurrent engineering methods, which decrease development time [2]. Both physics-based models and 3-D product modeling software give designers great creation and analysis abilities; however, both also share one potential limitation: a time-consuming need for detail.

Three-dimensional product modeling software and physics-based modeling can identify faulty designs and accurately predict performance; but, the accuracy of the estimations is limited by the design's level of detail, where a more-detailed input design will allow for better predictions of performance, and a model with less detail may include more assumptions and a larger degree of uncertainty. Unfortunately, the optimum level of detail is often not obvious to the designer until the latter stages of the design process. Additionally, as Mark Turner concluded in his report detailing the lessons learned from modeling the GE90 aircraft engine, though there is enough computing power to run detailed analyses, "…[computing power] is not as much of a bottleneck as the infrastructure for geometry definition, collection and grid generation for a given operating point" [3]. Defining the inputs for a complex design, whether it be an aircraft engine or a destroyer, can take a considerable amount of time, and therefore, most physics-based models and 3-D product modeling software are not conducive to effectively evaluating hundreds of designs during the conceptual design stage of a project. As part of research efforts sponsored by the Canadian Navy at the Aerospace Systems Design Laboratory, ASDL, at the Georgia Institute of Technology, another option was developed: integrate surrogate models with 3-D design visualization to produce a tool that gives near-instantaneous evaluation of a design that is easily defined.

This third option utilizes SHOP5, a legacy conceptual design tool developed by the Canadian Navy that applies naval architecture definitions and regressions of experimental and historical data to accurately and rapidly size and assess the performance of a conceptual design. As useful as SHOP5 can be, it requires the creation of an input file, and though modern computers can quickly write an input file, execute the program, and parse the output in about one third of a second, by using surrogate models, the analysis takes orders of magnitude less time. Calculating the formulas that comprise the surrogate models takes as little as a couple hundredths of a second, and when coupled with a graphical user interface makes evaluating a concept a process as easy as manipulating slide bars. Adding a visualization component allows a design to be modeled as a generalized ship that matches the input and output parameters, and this visualization works as a way to check the design to see if it "looks right" and possibly reject a design as infeasible. Another use is to notice trends and discern where to focus development efforts in order to affect the greatest improvement in performance; research by Chris McKesson highlighted how the use of relatively simple, parametric models can be used for exactly this purpose [4].

By matching the power of a legacy conceptual design tool augmented through surrogate models with an intuitive and responsive interface, it is possible to simplify the exploration of the design space, and by enabling rapid assessment may further enable the evaluation of great numbers of designs and improve the designers' mental model of the design space, effectively empowering them to make better decisions.

2.0 BODY

Before discussing the surrogate modeling process, it is necessary to describe and clarify the software used. SHOP5 stands for SHip OPtimization version 5, and is a legacy conceptual design tool originally developed in the late 1980s by James Colwell of Defense Research & Development Canada (DRDC.) It strives in analyzing batches of hundreds of designs; moreover, SHOP5 performs these analyses in a fraction of the time that other analyses tools require. However, when it comes to single designs, SHOP5's computational speed is jeopardized by the time required to write the input file and parse the output file. As a result, it can take practically the same amount of time to analyze either a single case or dozens of cases. Though in some applications this is acceptable or inconsequential, the process does induce a lag that significantly inhibits the performance of a real-time interface. Due to this lag, a different method must be employed to rapidly perform the required analyses.

One possible method is through the use of surrogate models. Surrogate models, being relatively simple formulas, have many benefits including being easy to implement in a program, calculating nearly instantaneously, and when properly created, accurately modeling a response, or output, throughout the entire bounds of a design space. The surrogate modeling process for this work is discussed in the latter sections of this paper. One important caveat is that before surrogate models can be created, data from which to create the models must be collected.

2.1 Design of Experiments

Before data collection can begin, the input and output variables to track must be selected. For this project the selected variables allow the user to control the design of the ship by altering characteristics in five separate categories: Characteristic Dimension; Geometry; Speeds; Combat, Power, and Complement; and Propulsion System. In total, twenty-four input variables are used. More variables can be included in the analysis, but for demonstration purposes, these were considered sufficient.

With the input variables selected, the Design of Experiments can now be formulated. A Design of Experiments, or DoE, is a systematic way of collecting the data required to make surrogate models. This method uses statistical methods described in work by Myers, Montgomery, and Anderson-Cook [5], and the objective is to limit the number of cases required to thoroughly explore the design space and make accurate models. Since only ordinal and continuous variables are suitable for DoEs, the engine configuration, a categorical variable, is not used in the DoE. For this project, the engine configuration is set to be CODOG, or COmbined Diesel Or Gas.

The remaining twenty-three variables are taken into consideration when choosing what DoE architecture to use. With such a large number of variables in use, a full factorial design is not feasible; for example, a three-level full factorial design requires 3^{23}, or 94,143,178,827, cases to be tested, and a face-centered central composite design is also infeasible since it requires $2^{23}+2(23)+1$, or 8,388,655, cases to be tested. If one assumes that it takes one-third of a second to run a case, it would take 32 days to run a face-centered central composite design and 994 years to run the full-factorial DoE. Since those lengths of time are practically beyond the reach of this project, a different tactic is used: Latin Hypercube / Hypercube designs.

The Latin Hypercube Design, like other space-filling designs, is beneficial because it allows the user to specify the number of cases. This design places the cases throughout the hyper and not just at the corners. Through use of the mathematical software MATLAB®, the DoE is tailored to reduce statistical correlation, an indication that there may be skewing in the results of the cases run and, most importantly, that models made from the data will be erroneous because the impact of the independent variables cannot be discerned independently. For this DoE, a Latin Hypercube with 17,250 cases was defined for the 23 variables; though this number appears to be high, it gives a sampling of cases that extend to the edges of the design space and minimizes the unexplored spaces.

In addition to the Latin Hypercube defined cases, there are 5700 random cases added to the DoE. These random cases are used to assess the fit of the surrogate model, but are not used to create the models, so any correlation between them is inconsequential.

With the entire DoE defined, a MATLAB® script is used to write the input file using input values from the DoE, send the input file to the SHOP5 executable, parse the output file, and save the output values to a data table. This completely automates the process of running a DoE; however, there is no simple way to track how many cases lead to failed analyses while the run is in progress. This is an issue because SHOP5 can reject a design for many reasons, and when it does so, it gives a coded message indicating the error, but otherwise produces no output for that design. The results from the initial DoE show that this happened 4826 times out of 17250 Latin Hypercube cases, or in 28% of the cases; this high percentage of failed cases renders the DoE results useless for meaningful surrogate modeling. Through the use of JMP®, a statistical and visual analytics software developed by the SAS Institute Inc., an investigation into the cause of the failed cases shows that most involved low full load displacement values; this is displayed below in Figure 1 where the darker shading indicates the distribution of the failed cases, and the lighter shading shows the distribution of displacement values for all Latin Hypercube cases.

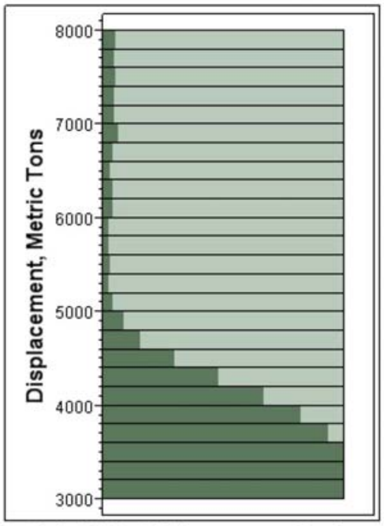


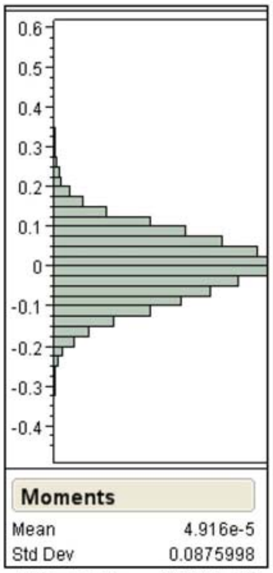
Figure 1. Initial DoE histogram. Darker shading shows where failed cases are.

Further investigation of error reports indicates that there is a conflict between low values of full load displacement and high values of combat systems weight. This conflict is often demonstrated by the inability to add fuel weight to the vessel, and hence the design has a predicted range of zero. This problem was solved by raising the minimum value of full load displacement, and after the new DoE was run, it was found that only 732 out of 17250 cases, or 4.24%, failed.

Though a small percentage of failed cases remains, a tradeoff between preventing failed cases and opening the design space is considered. For this project, a 4.24% failure rate is accepted to maintain the size of the design space and usefulness of the resultant tool. With the DoE data now collected, the surrogate modeling process can begin.

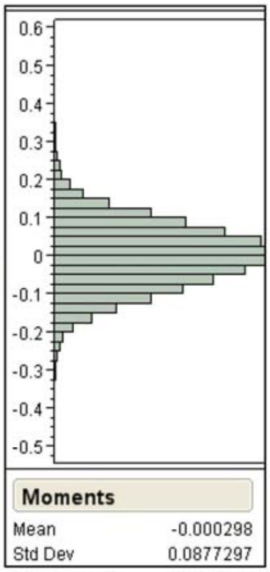
2.2 Surrogate Modeling

For this project, a combination of response surface equations and artificial neural networks are used to create the surrogate models for the Conceptual Ship Design Interface. Response surface equations used in this project take the form of linear, second order equations. By using general guidelines and processes developed by the ASDL, response surface equations are created with JMP®. Since these surrogate models are representations of other models, it is essential that the surrogate models accurately represent the behavior of the SHOP5 software throughout the design space. To establish how well a surrogate model actually models a process, 5 metrics are used: R^2 value, actual-by-predicted plots, residual-by-predicted plots, Model Fit Error (MFE,) and Model Representation Error (MRE.) Though all give important information about surrogate model performance, the most important are MRE and MFE. Model Fit Error measures the percent error between the prediction formula and the fitting data, and tells how well the model approximates the SHOP5 process at the fitting data points only. Model Representation Error tells how well the surrogate model represents the SHOP5 process throughout the design space by measuring the percent error between the prediction formula and all the data in the DoE. In practical terms, the 17250 Latin Hypercube cases are used to fit the model and determine the MFE, and the 5700 random cases are used with the Latin Hypercube cases to calculate the MRE. Surrogate models that accurately model the background process usually have percent error distributions that approximate normal distributions with a mean of zero and a standard distribution of 1. Figure 2 and Figure 3 show the MFE and MRE, respectively, for the surrogate model of draft; since the mean for each distribution is approximately zero and the standard deviation for each is significantly less than one, it is concluded that this is a very well-fitting surrogate model.



```
Mean        4.916e-5
Std Dev    0.0875998
```

Figure 2. MFE of the surrogate model for draft.



```
Mean       -0.000298
Std Dev    0.0877297
```

Figure 3. MRE of the surrogate model for draft.

Though many outputs, or responses, are successfully modeled with linear, second order response surface equations, some require a transformation to allow for a better fit. By modeling the logarithmic, or log, transformation of the response, it is observed that some outliers are reduced, and the surrogate model can better represent the response throughout the design space; moreover, though there is an additional calculation where the surrogate model value must be untransformed, modern computers are fast enough that the time required for this additional step is unnoticeable.

Logarithmic transformation improves the modeling of some responses, but there are a few that require an entirely different model.

In the case where linear, second order response surface equations prove to be inadequate for modeling some responses, artificial neural networks are tried. Artificial neural networks simulate the interaction between neurons and consist of a set of hidden nodes that perform the calculations that transform inputs to outputs. One of the strengths of artificial neural networks is their ability to accurately model non-linear responses, a useful trait for responses that are not modeled well with linear response-surface equations. The artificial neural networks are created through the use of a MATLAB® script based tool developed at the ASDL. This tool, BRAINN, was created by Carl Johnson and Jeff Schutte, and it allows for the semi-automated creation of artificial neural networks. During the creation of the neural nets, the MFE and MRE distributions are tracked, and different combinations of hidden nodes and training times are tried. For this project, a large range of number of hidden nodes is tested with short training times. Based on the best number of nodes, the range of number of hidden nodes is narrowed down, the training time is steadily increased, and the process is repeated until the best number of hidden nodes is found. This best number of hidden nodes is used for the artificial neural network, and the model performance is checked with the same metrics used to check the performance of response surface equations.

Artificial neural networks are used to model the remaining responses, and the investigation of the performance of these surrogate models shows that they accurately predict the responses they model. With accurate surrogate models now ready to be used, work on integrating the models into a useful interface can begin.

2.3 The Conceptual Ship Design Interface

The surrogate models are used in the background of a graphical user interface, the Conceptual Ship Design Interface, or CSDI; this interface is used to allow a user to rapidly and easily define a design and receive outputs in real-time. The CSDI is written in JMP® scripting language, and it is a relatively straightforward process of linking the surrogate models into the scripting framework. The challenge comes from creating a parametric method to define 3-D ship geometry. To create the parametric model, the ship was broken into three parts: the hull below the waterline, the hull above the waterline, and the superstructure.

The first section defined is the hull below the waterline. Three naval architecture coefficients are used to describe the hull: block, prismatic, and midship coefficients. The midship coefficient is the ratio of the hull cross-section area to a square area defined by the beam at the waterline and the draft; additionally, the midship coefficient is the resultant of the block coefficient divided by the prismatic coefficient. For the interface, the cross-section shape of the hull is a function of the beam, draft, and midship coefficient. Figure 4 shows the parameters used to define the underwater panels of the hull.

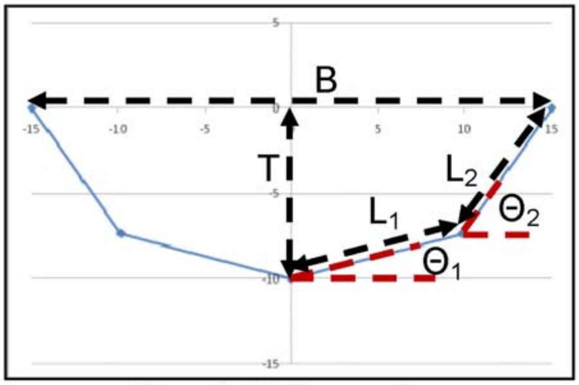


Figure 4. Hull parameters.

In Figure 4, B is the beam of the vessel, T is the draft of the vessel, L_1 and L_2 are panel lengths, and θ_1 and θ_2 are shaping parameters. By analyzing the geometry, the panel lengths can be solved for in terms of the other.

$$L_1 = \frac{-(2T\cos\theta_2 - B\sin\theta_2)}{2(\cos\theta_1\sin\theta_2 - \sin\theta_1\cos\theta_2)} \quad (1)$$

$$L_2 = \frac{(2T\cos\theta_1 - B\sin\theta_1)}{2(\cos\theta_1\sin\theta_2 - \sin\theta_1\cos\theta_2)} \quad (2)$$

Using Eq. (1) and Eq. (2,) it is possible to define the midship coefficient, or C_m, in terms of the panel lengths, shaping parameters, beam, and draft of the hull. For the CSDI, θ_1 is set at a constant five degrees, and an algorithm iterates through values of θ_2 until the calculated value of the midship coefficient matches the specified value. The shaded areas in Figure 5 and Figure 6 show the different hull forms that can be created by using this method; in both figures, θ_1 is five degrees, and θ_2 is set to thirty five degrees and 85 degrees, respectively. For reference, the midship coefficient for the cross-section in Figure 5 is 0.523, and it is 0.912 for the cross-section in Figure 6.

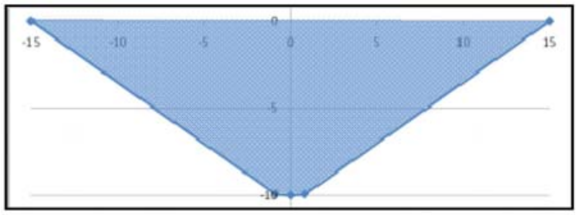


Figure 5. Hull cross-section with low midship coefficient.

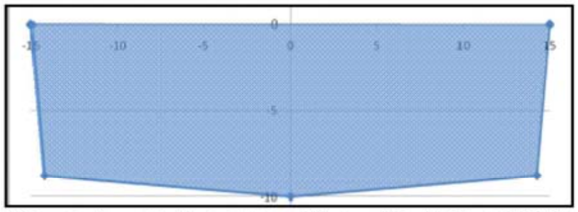


Figure 6. Hull cross-section with high midship coefficient.

The cross-section defined by the process detailed above is applied along the length of the ship, and since the panels are determined as a function of the beam, the cross-section is automatically sized for the local beam. The local beam is defined as a fraction of the largest, or in this case midship, beam. A theoretical hull design, inspired by the design of international frigates, is drawn up, and from this design, the beam fractions are measured. These beam fractions are then used to fit a model through the use of JMP®, and the resulting equation is used to define the local beam length as a function of the fractional length along the hull at the waterline. A similar process is used to define the local beams of the hull at the top of the hull, but the theoretical hull shape is modified to account for flaring. At the current version, the CSDI does not allow the user to modify the shape of the hull; the hull is merely scaled up or down with the ship length, beam, and draft.

These local beam functions are used to define the hull panels below and above the waterline; however, there is a special consideration for the hull panels above the waterline. To allow for the lowering of the flight deck, some additional steps are required. First, the area of the flight deck is approximated as a trapezoid defined by the port and starboard edges of the flight deck and the forward and aft local beams of the flight deck. Since the hull along the flight deck and above the waterline is defined as a single panel, the trapezoidal approximation works well. With the flight deck area calculated, the maximum allowed amount of lowering is calculated. In order to account for performance changes due to the lowering of the flight deck, a hull volume margin is one of the variables used in the DoE. However, the maximum hull volume margin accounted could allow for different amounts of flight deck lowering based on the size of the flight deck. As a result, maximum amount of lowering must be recalculated whenever the flight deck length or ship size changes. Once the maximum is calculated, the input value of lowering is checked, and if it is less than the maximum allowed amount, the flight deck is lowered by that amount. Once the flight deck is lowered, however, it is no longer correct to use the top local beam function for the top of the lowered hull. To prevent rendering problems, the local beam values of the flight deck are interpolated between the top local beam function value and waterline local beam function. Though doing so would be noticeable in the resultant visualization, there is actually little difference in the local beam created by this interpolation, and any changes to the area of the flight deck due to the interpolation are not noted in future analyses.

With the hull completely defined, the last section of the ship to be modeled is the superstructure. The superstructure cross-section is trapezoidal, and though the user can input the superstructure height in the interface, the front surface area remains constant with respect to the height, though it can change as the local beam changes; this is to account for how SHOP5 models the front area of the superstructure. Once the front surface area is calculated, the length of the superstructure is calculated from the surrogate models. If this length is greater than sixty percent of the length of the ship at the waterline, the visualization program overrides this calculation and sets the superstructure length to sixty percent of the length of the ship at the waterline in order to prevent rendering problems. With these parameters now defined, the modeling of the superstructure begins.

In the visualization, the superstructure is modeled ‘backwards,’ that is, from the back to the front. The back of the superstructure is defined to start at the forward end of the flight deck. This is done in consideration of the hangar, which must interact with the flight deck in a realistic design. Panels for the superstructure are then defined forwards until the superstructure reaches the previously defined length. On each end of the superstructure, a slight slope is added; to improve the aesthetics of the design. The superstructure is the last part of the ship to be modeled, and an example ship, as visualized in the CSDI, is presented as Figure 7.

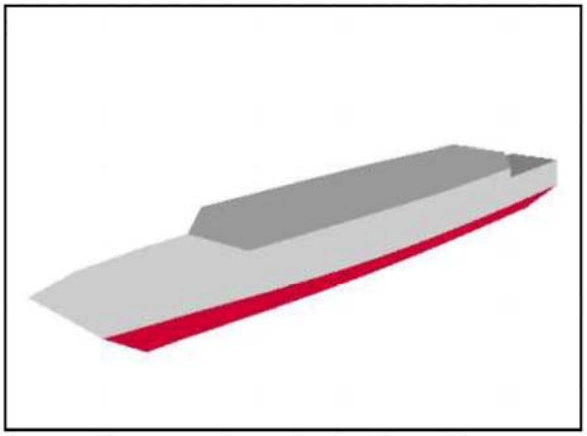
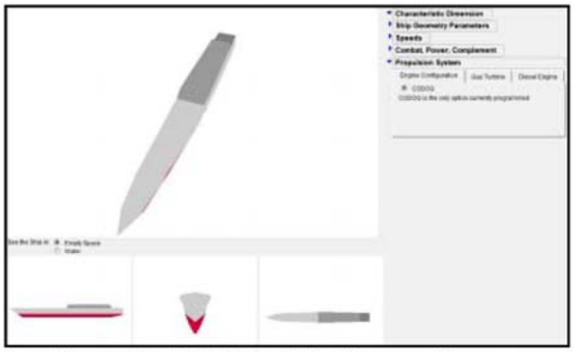


Figure 7. Example visualized ship.

The visualization is integrated into CSDI, and instead of a single, isometric view, the user is presented with an isometric view, a side view, a front view, and a top view. Each view can be independently adjusted, enabling the user to move around the ship and inspect the design from any exterior angle. Figure 8 shows these visualization windows and the column of input categories.



- Characteristic Dimension
- Ship Geometry Parameters
- Speeds
- Combat, Power, Complement
- Propulsion System

● COGOG

COGOG is the only option currently programmed

See the Ship in: ● Empty Space ○ Water

Figure 8. CSDI visualizations and inputs.

3.0 DISCUSSION

Throughout the process of this project, some important observations are made. During the creation of the design of experiments, investigating the interplay between input variables developed into a strategy for making a better DoE. Additionally, in the surrogate modeling process it is observed that no single type of model works for every response; it takes a combination of response surface equations and artificial neural networks to adequately model the responses. Finally, in the course of creating the visualization, it is determined that the cross-section of the hull can be defined by as little as one parameter: midship coefficient. These important observations notwithstanding, the CSDI is the most important outcome.

With the creation of the Conceptual Ship Design Interface, this project enables the rapid investigation of the design space for ship design. The CSDI gives the user the power to discover trends throughout the ship design space, and through the visualization methods developed for this project, the CSDI shows the user what the conceptual ship could look like, thus allowing the user to rapidly reject faulty designs. Though the CSDI can be useful, it still has limits; the full load displacement limits restrict the use of the CSDI to defining heavy frigate and destroyer designs. This limit is the focus of following efforts, and it is expected that by creating different classes of ships with different full load displacement ranges will remove this limitation.

4.0 CONCLUSIONS

By applying advanced aerospace systems engineering methods to the field of naval architecture, this project developed a new approach to conceptual ship design. The use of surrogate models, instead of linking to a program, allows for real-time feedback as the user specifies the design. The final product of the project, the Conceptual Ship Design Interface, uses these models and presents the user a graphical interface that is capable of determining feasibility and estimating performance. As a result, the CSDI enables rapid assessment of designs, and with this capability, it empowers designers to make better decisions throughout the design process.

5.0 REFERENCES

[1] Jiang, C., Forstell, B., Lavis, D., and Ritter, O., "Ship hull and machinery optimization using physics-based design software," *Marine*

Technology and SNAME News, Vol. 39, No. 2, Apr. 2002, pp. 109-117.

[2] Baum, S. J., and Ramakrishnan, R., "Applying 3D product modeling technology to shipbuilding," *Marine Technology and SNAME News*, Vol. 34, No. 1, Jan. 1997, pp. 56-65.

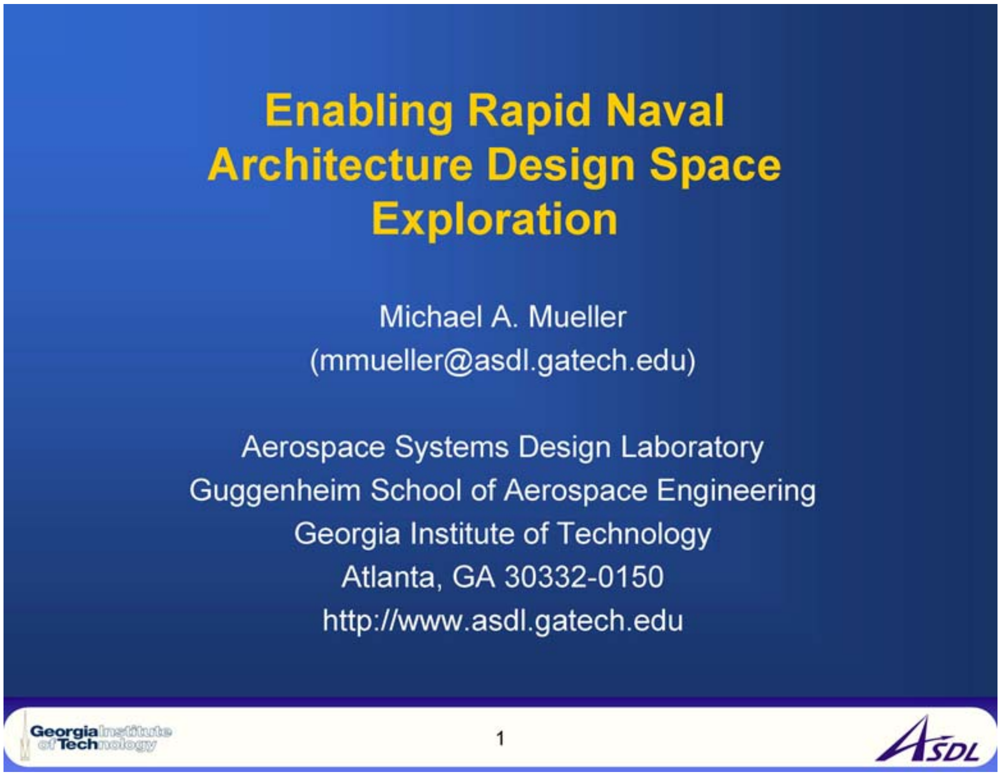
[3] Turner, M., "Lessons Learned from the GE90 3D Full Engine Simulations," *48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition*, AIAA 2010-1606-796, AIAA, Washington, DC, 2010, pp. 1-16.

[4] McKesson, C., "The Utility of Very Simple Models for Very Complex Systems," *2010 International Simulation Multiconference* [CD-ROM], Ottawa, Ontario, 12-14 Jul. 2010, pp. 181-187.

[5] Myers, R. H., Montgomery, D. C., Anderson-Cook, C. M., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 3rd ed., John Wiley & Sons, Inc., Hoboken, New Jersey, 2009, Chaps. 7, 8.

6.0 ACKNOWLEDGMENTS

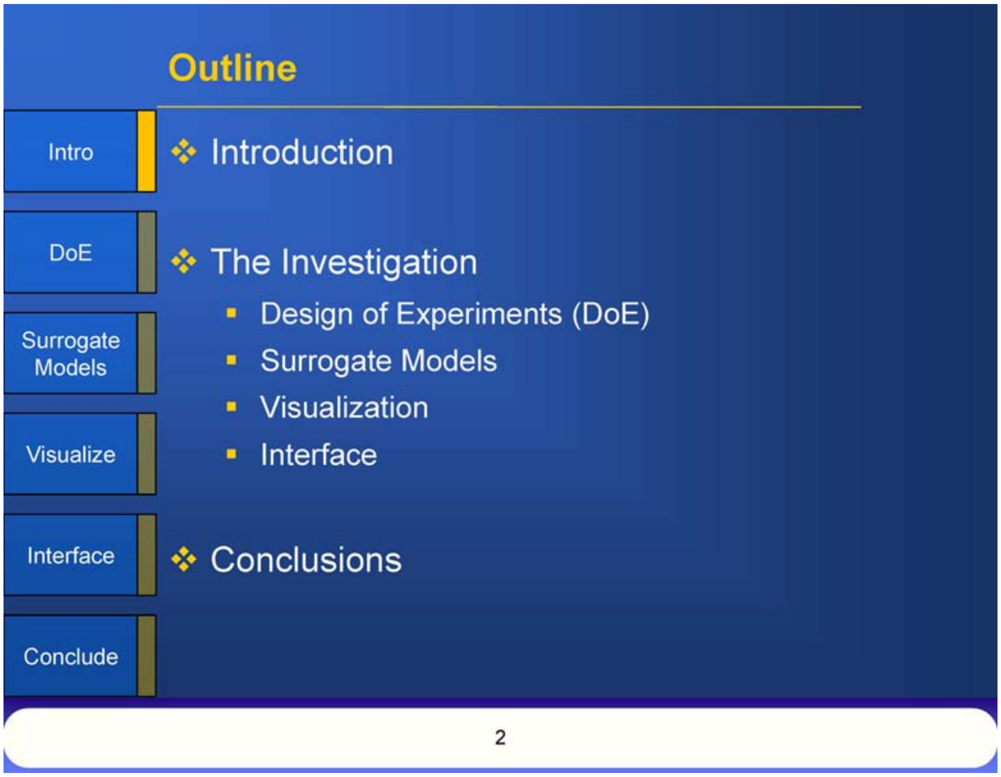
The authors thank the Canadian Navy, particularly LCdr Jacques Olivier and Mr. Sean Tobin, for their support of this project. The authors also thank Dr. James Colwell from Defence Research & Development Canada for sharing his SHOP5 expertise with us.



Enabling Rapid Naval Architecture Design Space Exploration

Michael A. Mueller
(mmueller@asdl.gatech.edu)

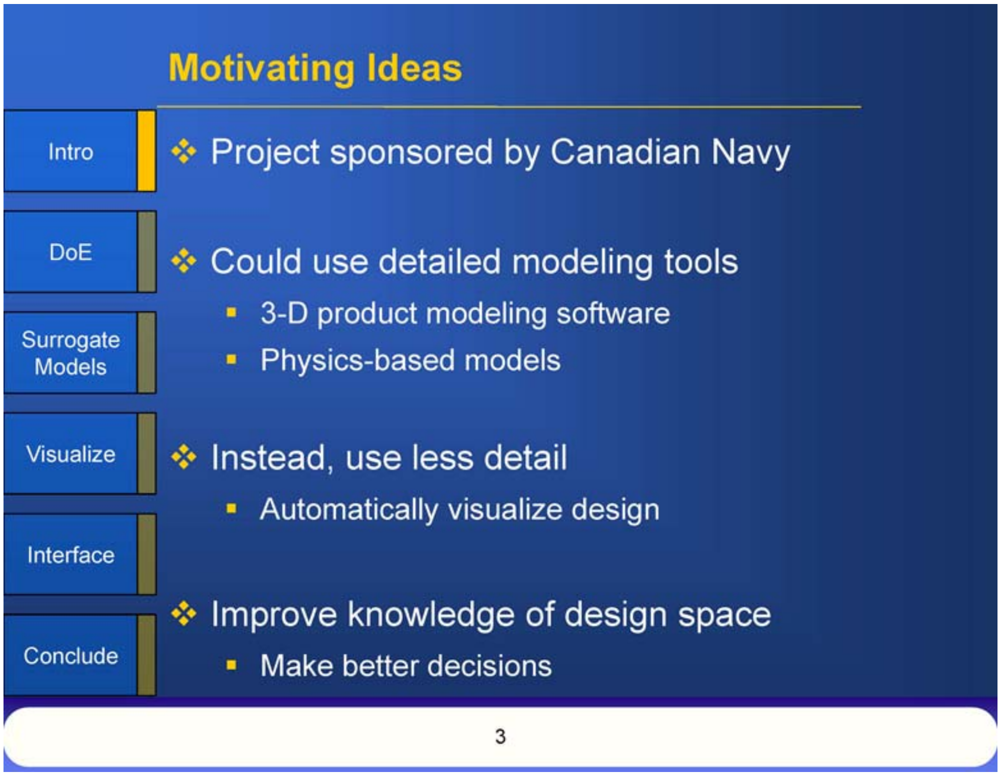
Aerospace Systems Design Laboratory
Guggenheim School of Aerospace Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0150
http://www.asdl.gatech.edu



Outline

Intro
DoE
Surrogate Models
Visualize
Interface
Conclude

- Introduction
- The Investigation
  - Design of Experiments (DoE)
  - Surrogate Models
  - Visualization
  - Interface
- Conclusions

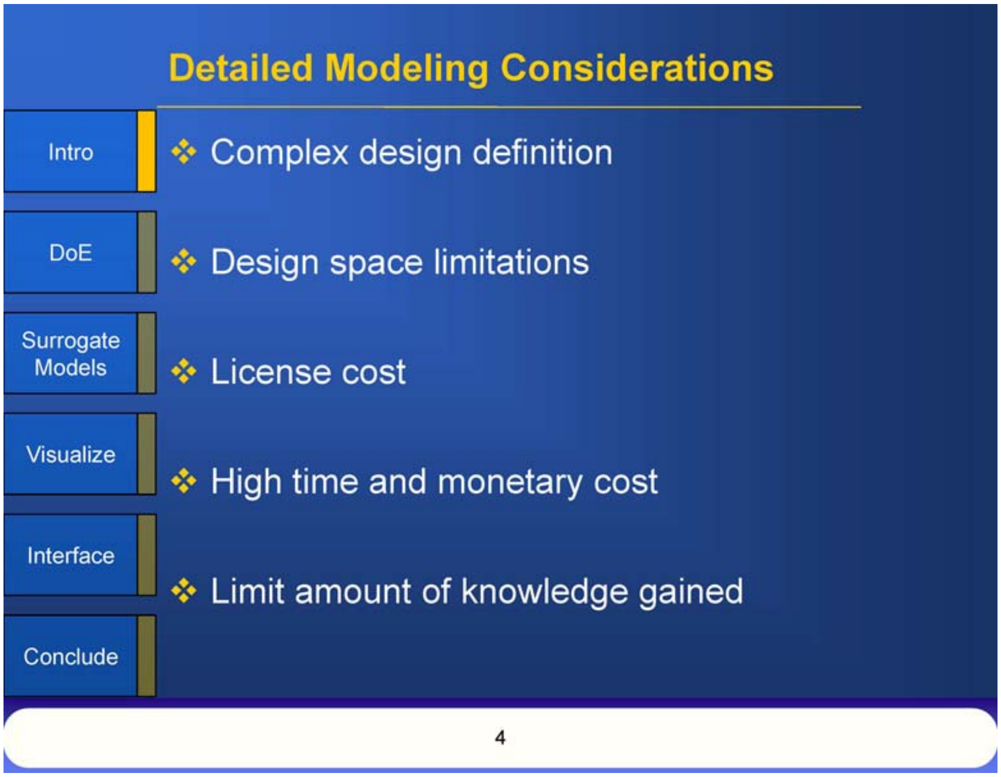


Motivating Ideas

Intro
DoE
Surrogate Models
Visualize
Interface
Conclude

- Project sponsored by Canadian Navy
- Could use detailed modeling tools
  - 3-D product modeling software
  - Physics-based models
- Instead, use less detail
  - Automatically visualize design
- Improve knowledge of design space
  - Make better decisions

3

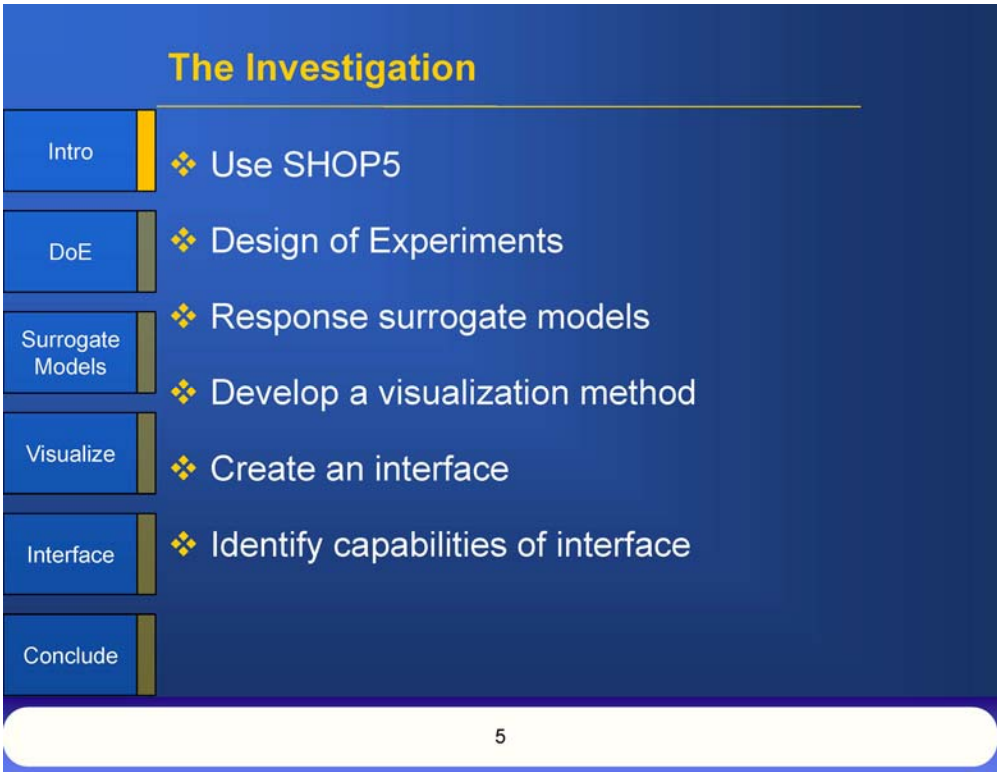


Detailed Modeling Considerations

Intro
DoE
Surrogate Models
Visualize
Interface
Conclude

- Complex design definition
- Design space limitations
- License cost
- High time and monetary cost
- Limit amount of knowledge gained

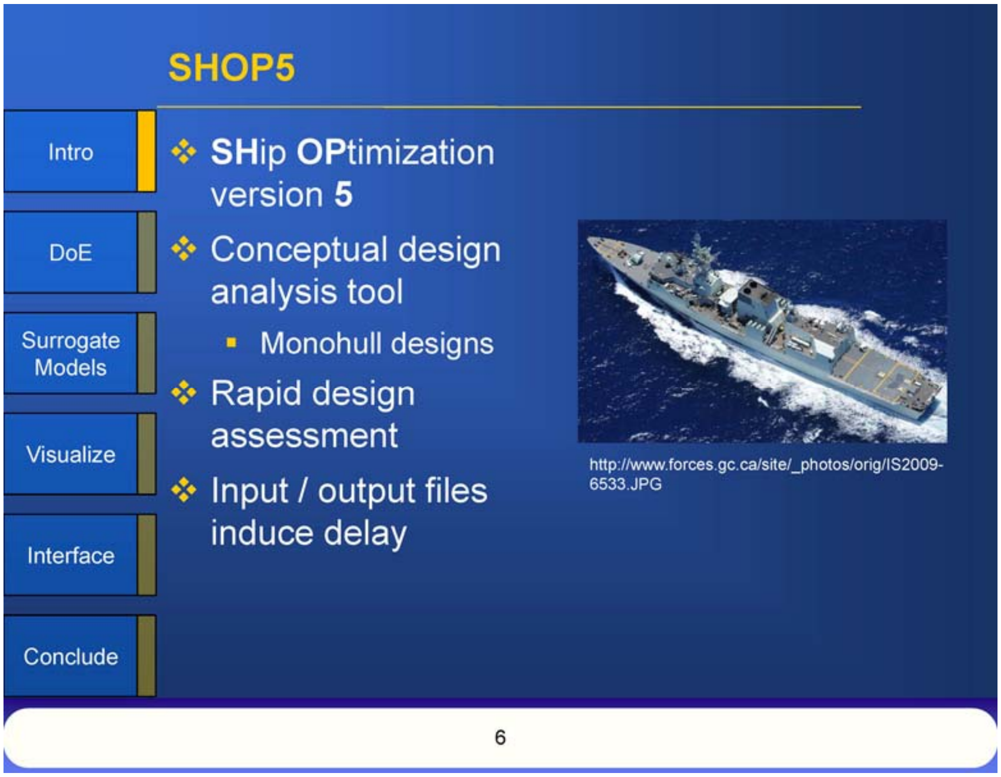
4


Intro
DoE
Surrogate Models
Visualize
Interface
Conclude

The Investigation

- Use SHOP5
- Design of Experiments
- Response surrogate models
- Develop a visualization method
- Create an interface
- Identify capabilities of interface

5

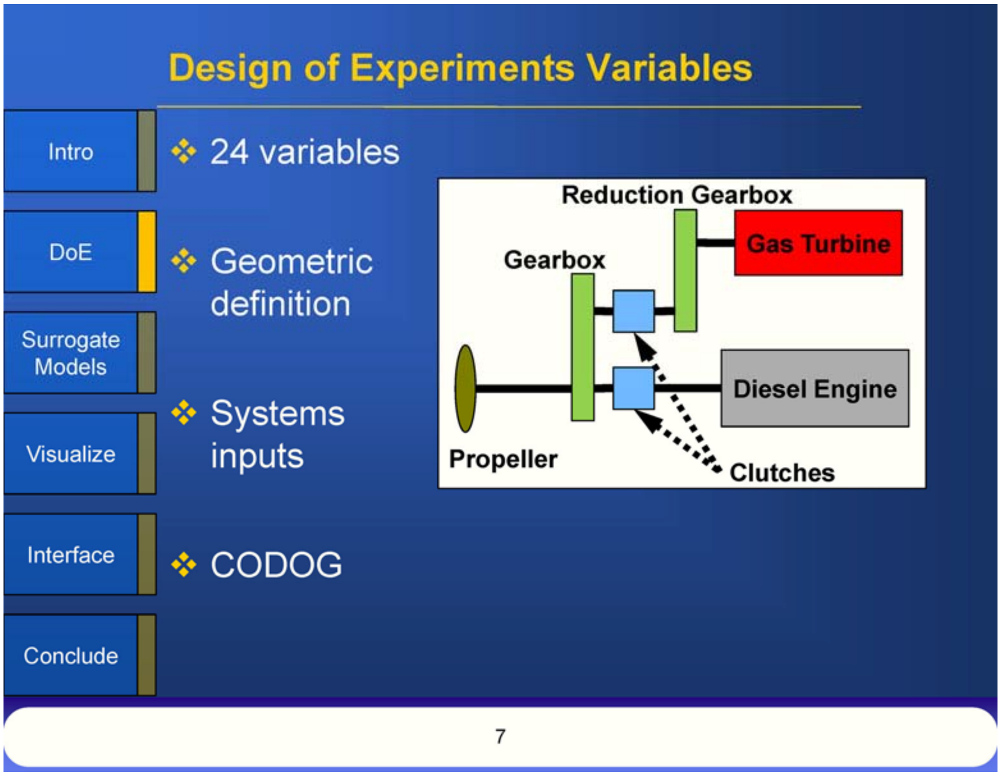

Intro
DoE
Surrogate Models
Visualize
Interface
Conclude

SHOP5

- **SH**ip **OP**timization version **5**
- Conceptual design analysis tool
  - Monohull designs
- Rapid design assessment
- Input / output files induce delay

http://www.forces.gc.ca/site/_photos/orig/IS2009-6533.JPG

6

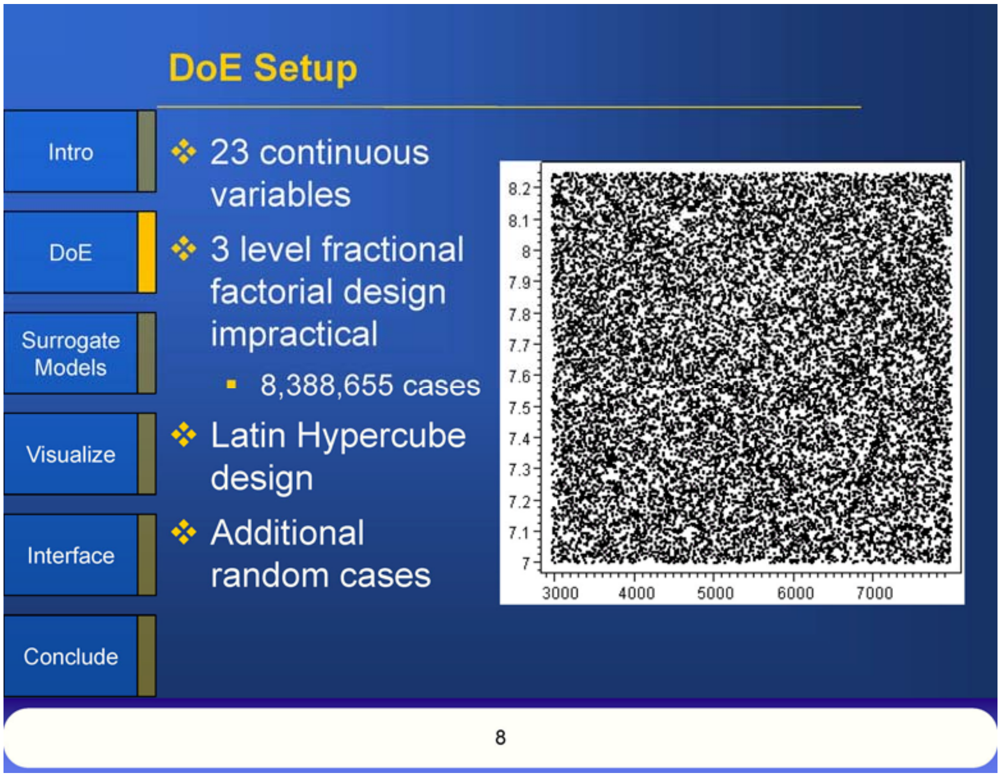


Design of Experiments Variables

Intro
DoE
Surrogate Models
Visualize
Interface
Conclude

- 24 variables
- Geometric definition
- Systems inputs
- CODOG

Reduction Gearbox
Gas Turbine
Gearbox
Diesel Engine
Propeller
Clutches

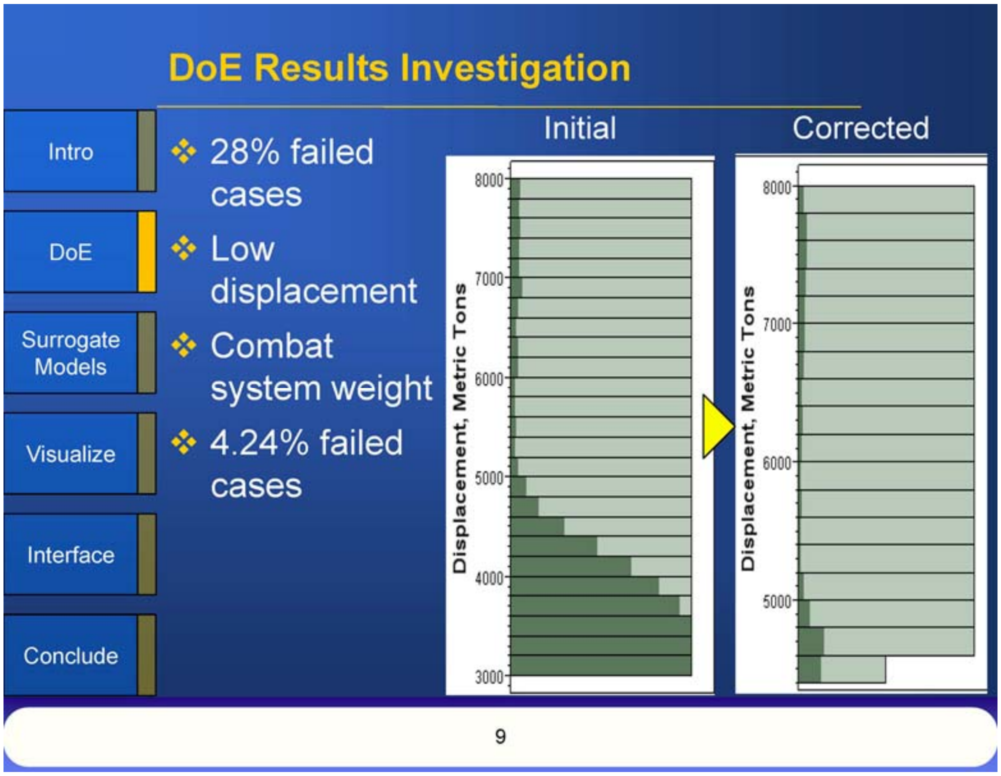


DoE Setup

Intro
DoE
Surrogate Models
Visualize
Interface
Conclude

- 23 continuous variables
- 3 level fractional factorial design impractical
  - 8,388,655 cases
- Latin Hypercube design
- Additional random cases

8.2
8.1
8
7.9
7.8
7.7
7.6
7.5
7.4
7.3
7.2
7.1
7
3000
4000
5000
6000
7000

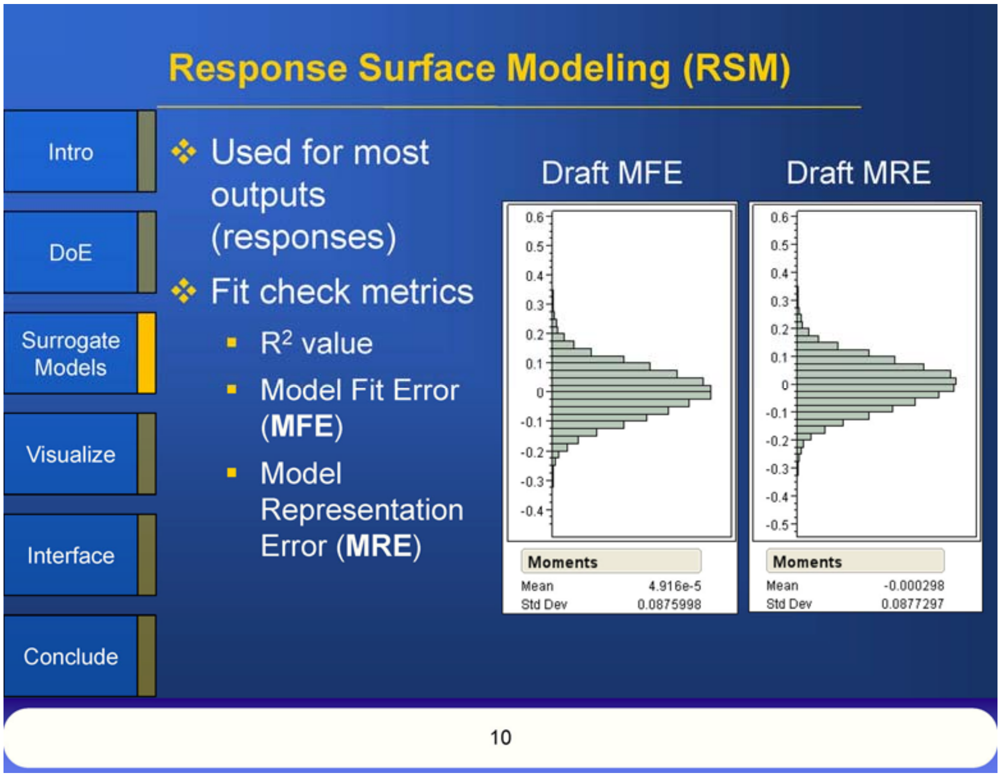


DoE Results Investigation

- Intro
- DoE
- Surrogate Models
- Visualize
- Interface
- Conclude

- 28% failed cases
- Low displacement
- Combat system weight
- 4.24% failed cases

9



Response Surface Modeling (RSM)

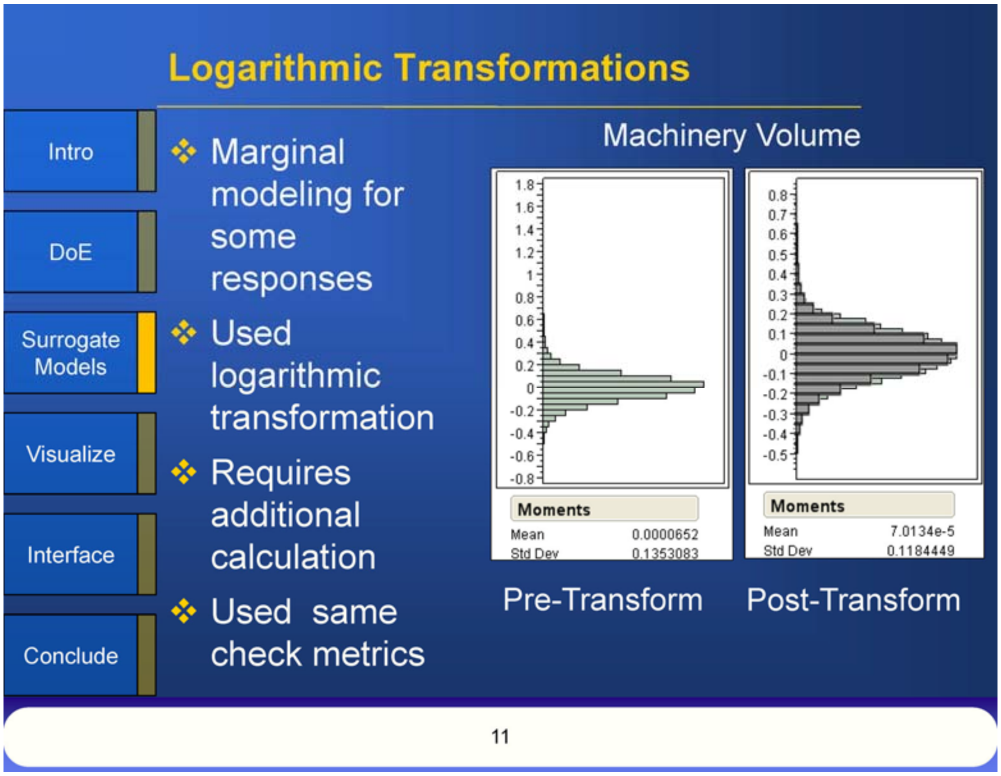
- Intro
- DoE
- Surrogate Models
- Visualize
- Interface
- Conclude

- Used for most outputs (responses)
- Fit check metrics
  - R^2 value
  - Model Fit Error (**MFE**)
  - Model Representation Error (**MRE**)

Moments	
Mean	4.916e-5
Std Dev	0.0875998

Moments	
Mean	-0.000298
Std Dev	0.0877297

10



Logarithmic Transformations

Intro
DoE
Surrogate Models
Visualize
Interface
Conclude

- Marginal modeling for some responses
- Used logarithmic transformation
- Requires additional calculation
- Used same check metrics

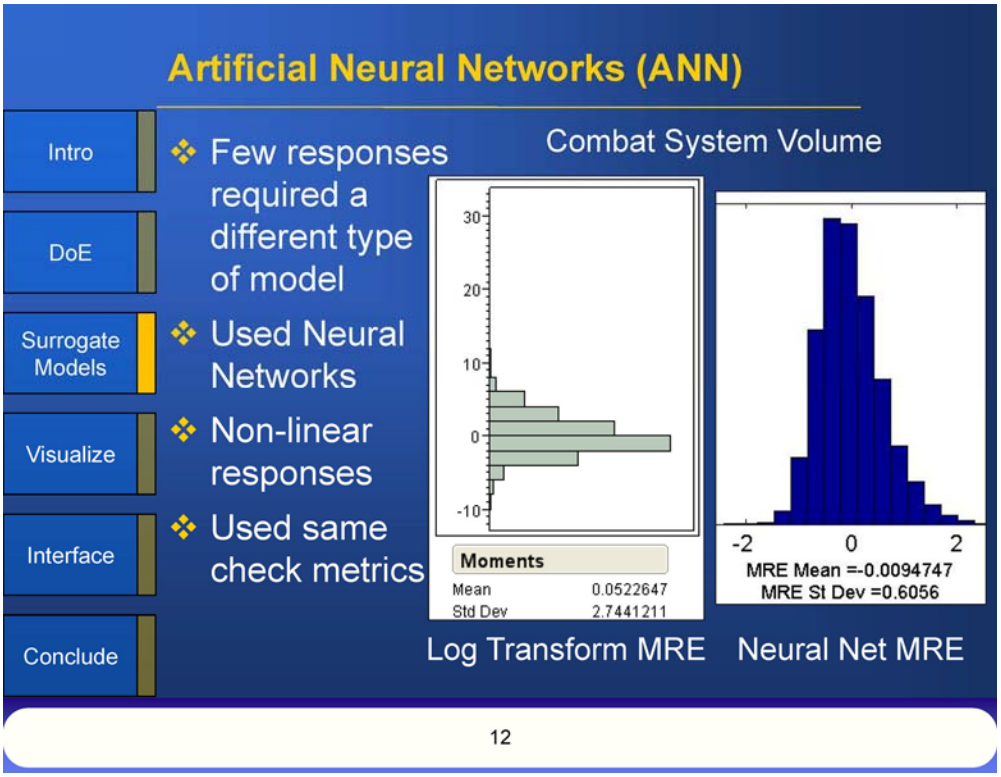
Machinery Volume

Moments	
Mean	0.0000652
Std Dev	0.1353083

Moments	
Mean	7.0134e-5
Std Dev	0.1184449

Pre-Transform Post-Transform

11



Artificial Neural Networks (ANN)

Intro
DoE
Surrogate Models
Visualize
Interface
Conclude

- Few responses required a different type of model
- Used Neural Networks
- Non-linear responses
- Used same check metrics

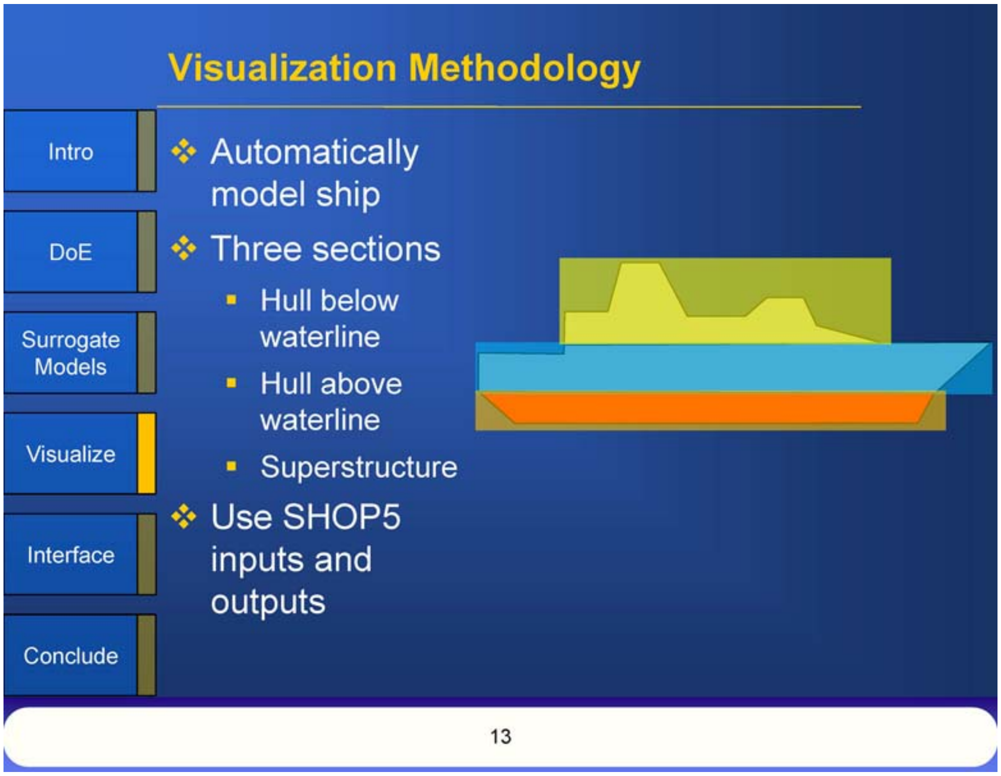
Combat System Volume

Moments	
Mean	0.0522647
Std Dev	2.7441211

MRE Mean =-0.0094747
MRE St Dev =0.6056

Log Transform MRE Neural Net MRE

12

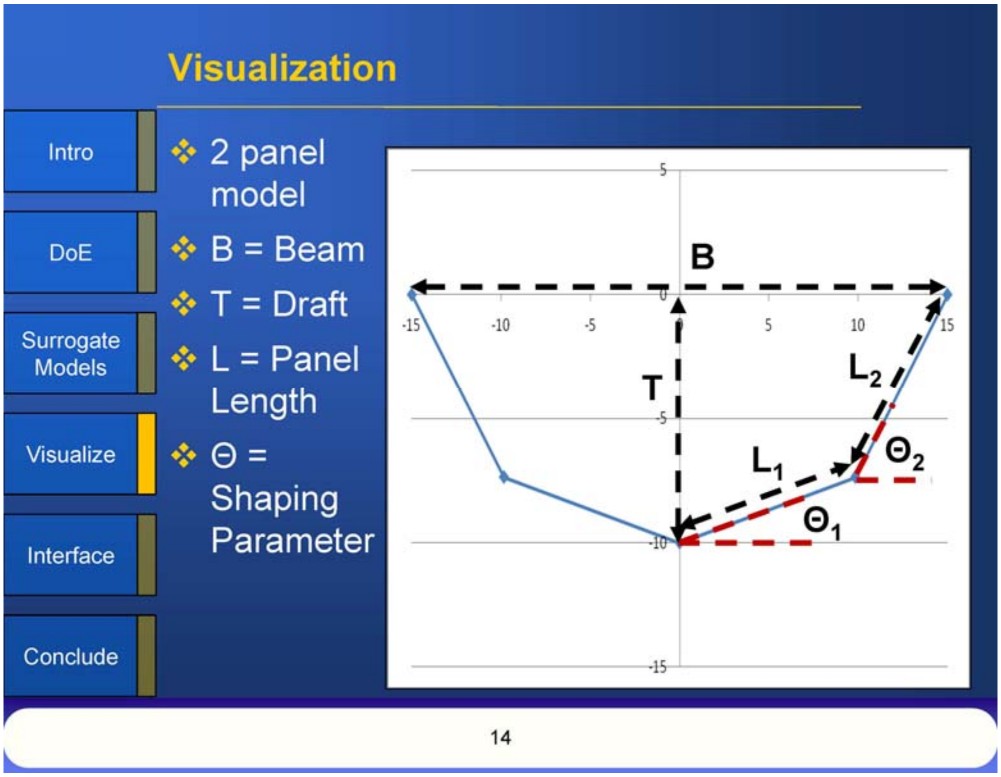


Visualization Methodology

- Intro
- DoE
- Surrogate Models
- Visualize
- Interface
- Conclude

- Automatically model ship
- Three sections
  - Hull below waterline
  - Hull above waterline
  - Superstructure
- Use SHOP5 inputs and outputs

13

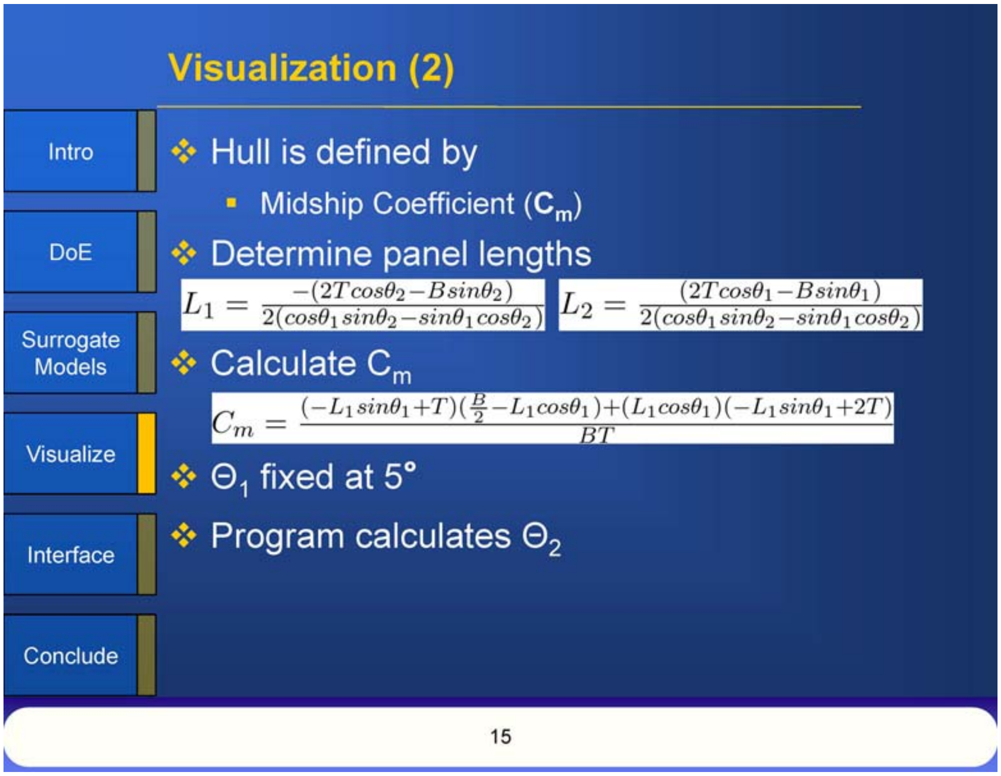


Visualization

- Intro
- DoE
- Surrogate Models
- Visualize
- Interface
- Conclude

- 2 panel model
- B = Beam
- T = Draft
- L = Panel Length
- Θ = Shaping Parameter

14



Visualization (2)

- Intro
- DoE
- Surrogate Models
- Visualize
- Interface
- Conclude

- Hull is defined by
  - Midship Coefficient (**C**$_m$)
- Determine panel lengths

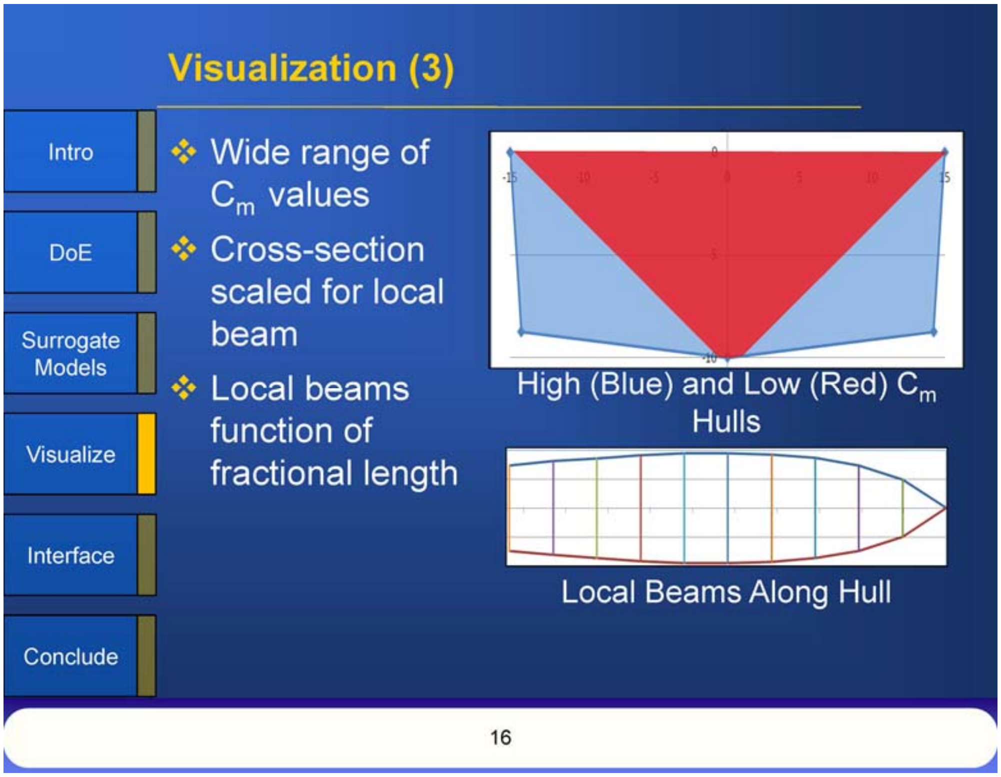
$$L_1 = \frac{-(2Tcos\theta_2 - Bsin\theta_2)}{2(cos\theta_1 sin\theta_2 - sin\theta_1 cos\theta_2)} \quad L_2 = \frac{(2Tcos\theta_1 - Bsin\theta_1)}{2(cos\theta_1 sin\theta_2 - sin\theta_1 cos\theta_2)}$$

- Calculate C$_m$

$$C_m = \frac{(-L_1 sin\theta_1 + T)(\frac{B}{2} - L_1 cos\theta_1) + (L_1 cos\theta_1)(-L_1 sin\theta_1 + 2T)}{BT}$$

- Θ$_1$ fixed at 5°
- Program calculates Θ$_2$

15



Visualization (3)

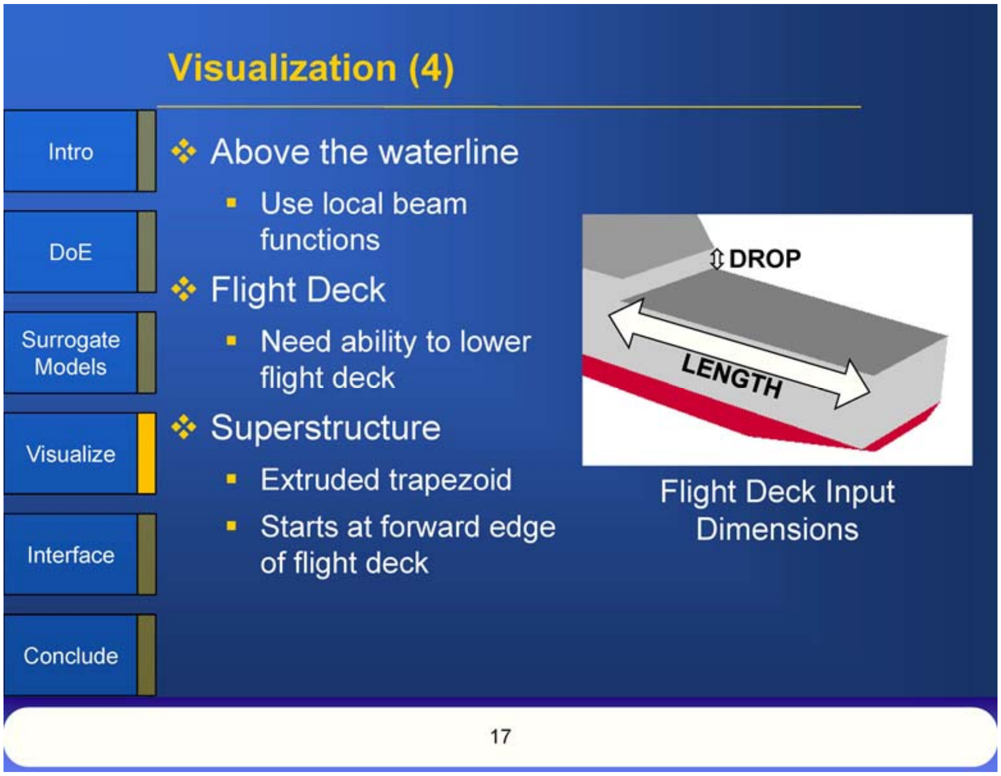
- Intro
- DoE
- Surrogate Models
- Visualize
- Interface
- Conclude

- Wide range of C$_m$ values
- Cross-section scaled for local beam
- Local beams function of fractional length

High (Blue) and Low (Red) C$_m$ Hulls

Local Beams Along Hull

16



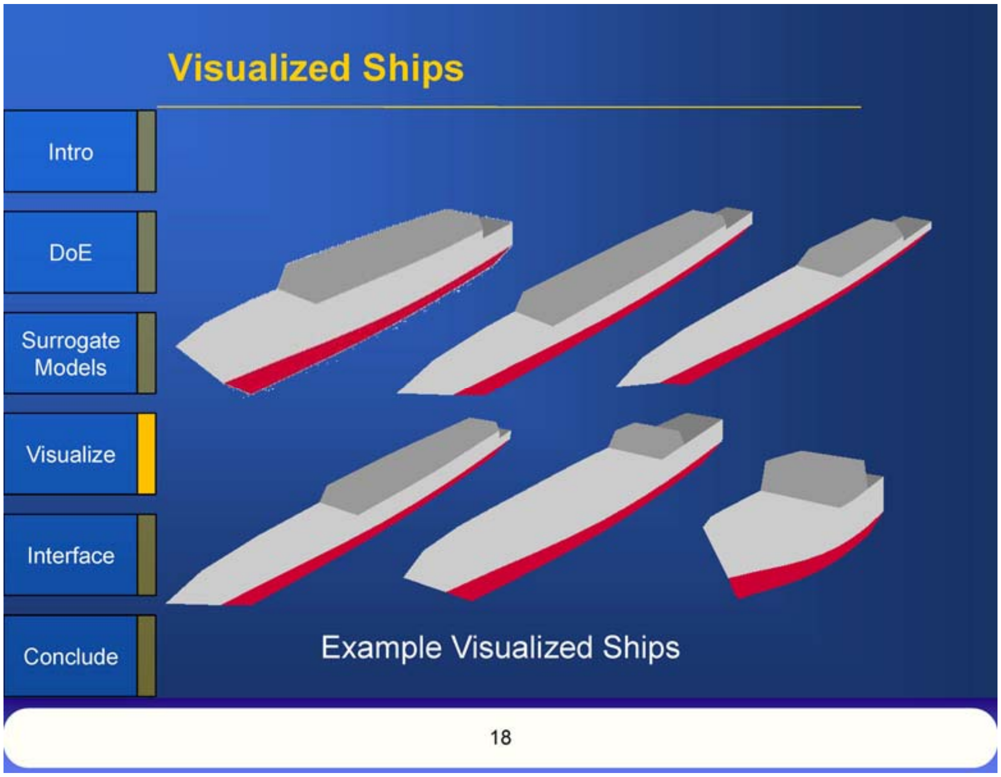
Visualization (4)

- Intro
- DoE
- Surrogate Models
- Visualize
- Interface
- Conclude

- Above the waterline
  - Use local beam functions
- Flight Deck
  - Need ability to lower flight deck
- Superstructure
  - Extruded trapezoid
  - Starts at forward edge of flight deck

Flight Deck Input Dimensions

17



Visualized Ships

- Intro
- DoE
- Surrogate Models
- Visualize
- Interface
- Conclude

Example Visualized Ships

18

Benefits of the Approach

Intro

DoE

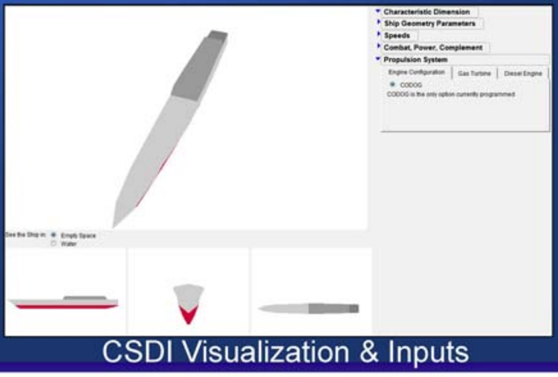
Surrogate Models

Visualize

Interface

Conclude

- Fast design definition
- Instantaneous feedback
- Visual interaction with design
- Discover trends with much less effort



● CODOG

CODOG is the only option currently programmed

See the Ship in: ● Empty Space ○ Water

CSDI Visualization & Inputs

19

Future Work

Intro

DoE

Surrogate Models

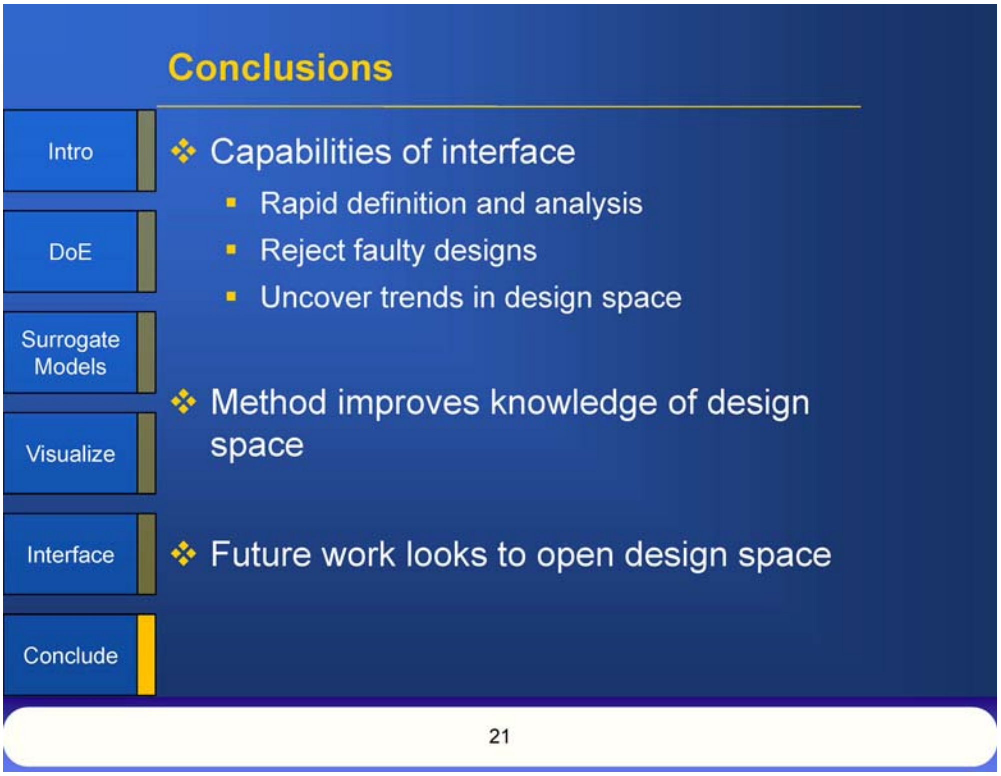
Visualize

Interface

Conclude

- Improve CSDI capabilities
  - Displacement
  - Engine configuration
- Transition to more detailed models
- Determine cost/capability tradeoffs

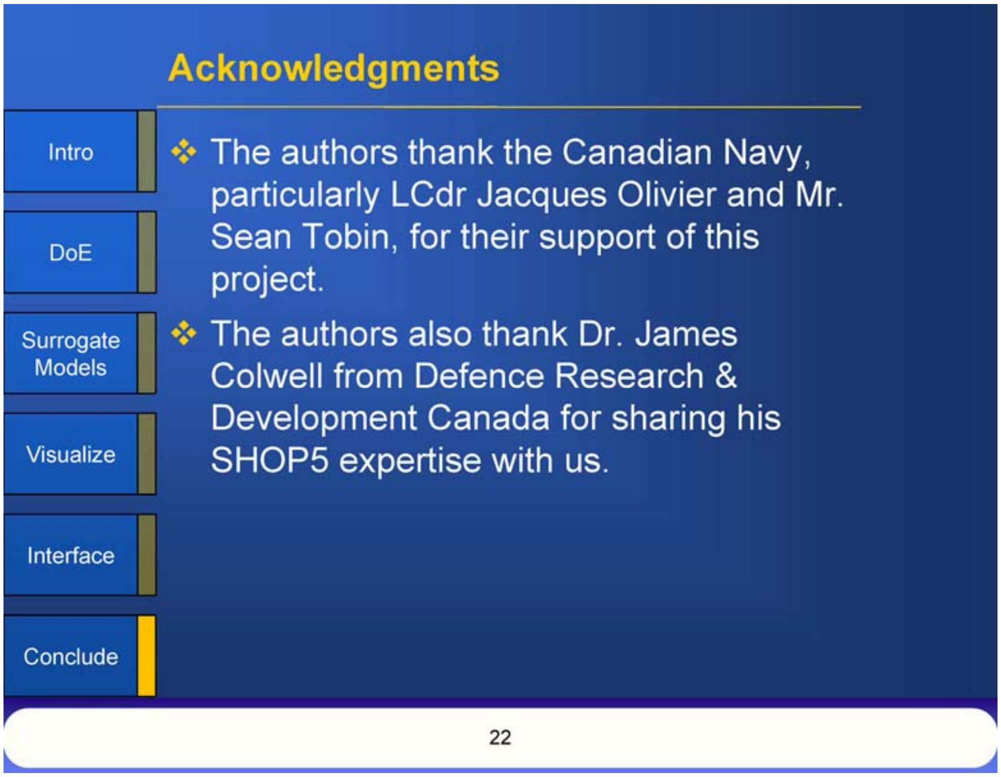
20


Intro
DoE
Surrogate Models
Visualize
Interface
Conclude

Conclusions

- Capabilities of interface
  - Rapid definition and analysis
  - Reject faulty designs
  - Uncover trends in design space
- Method improves knowledge of design space
- Future work looks to open design space

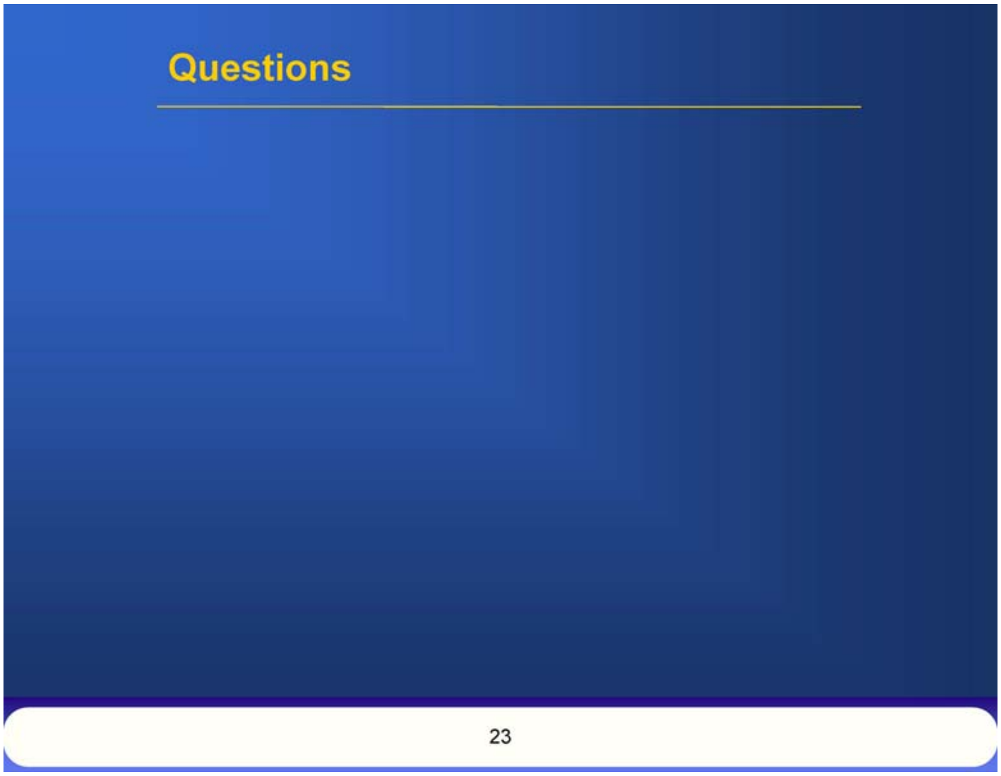
21


Intro
DoE
Surrogate Models
Visualize
Interface
Conclude

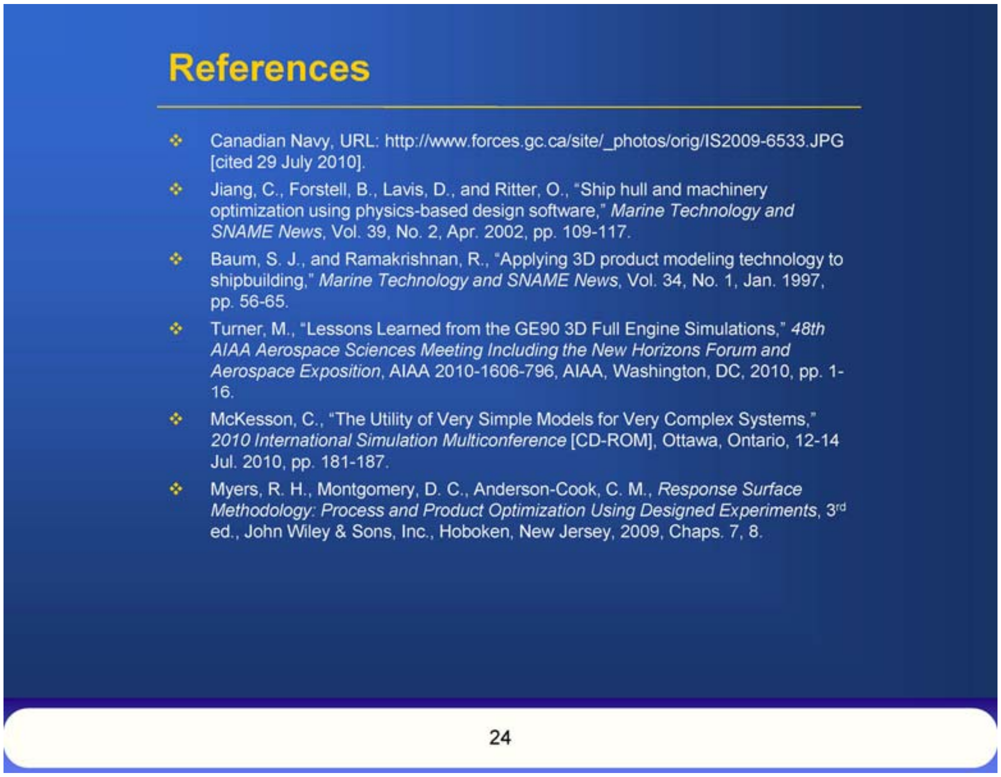
Acknowledgments

- The authors thank the Canadian Navy, particularly LCdr Jacques Olivier and Mr. Sean Tobin, for their support of this project.
- The authors also thank Dr. James Colwell from Defence Research & Development Canada for sharing his SHOP5 expertise with us.

22



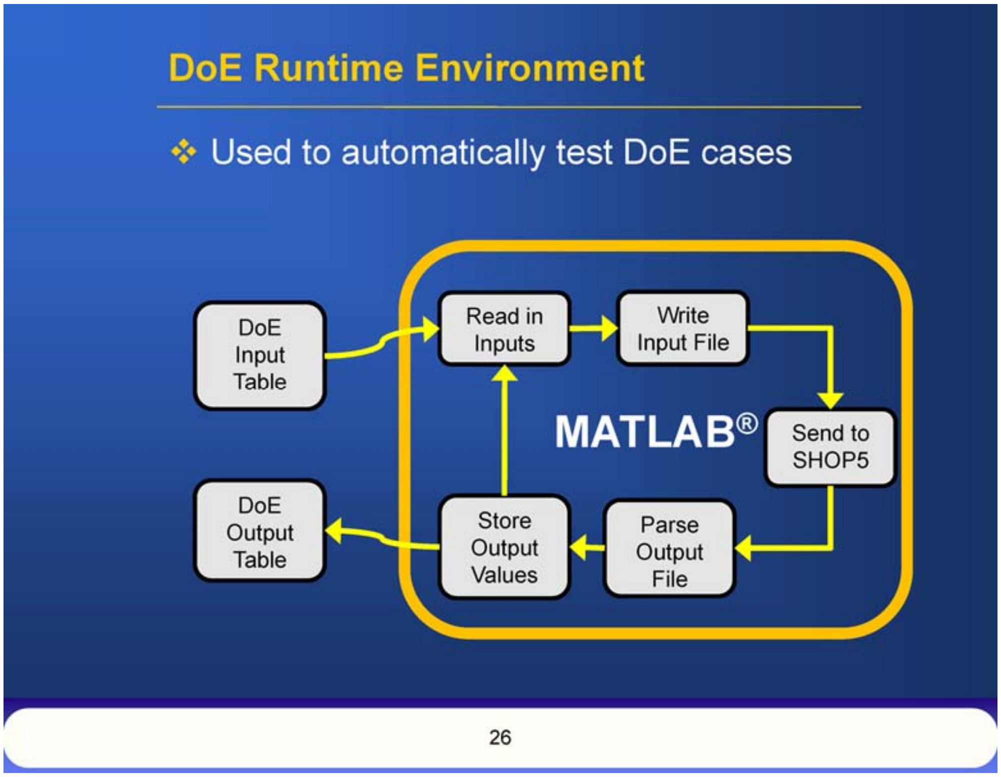
Questions



References

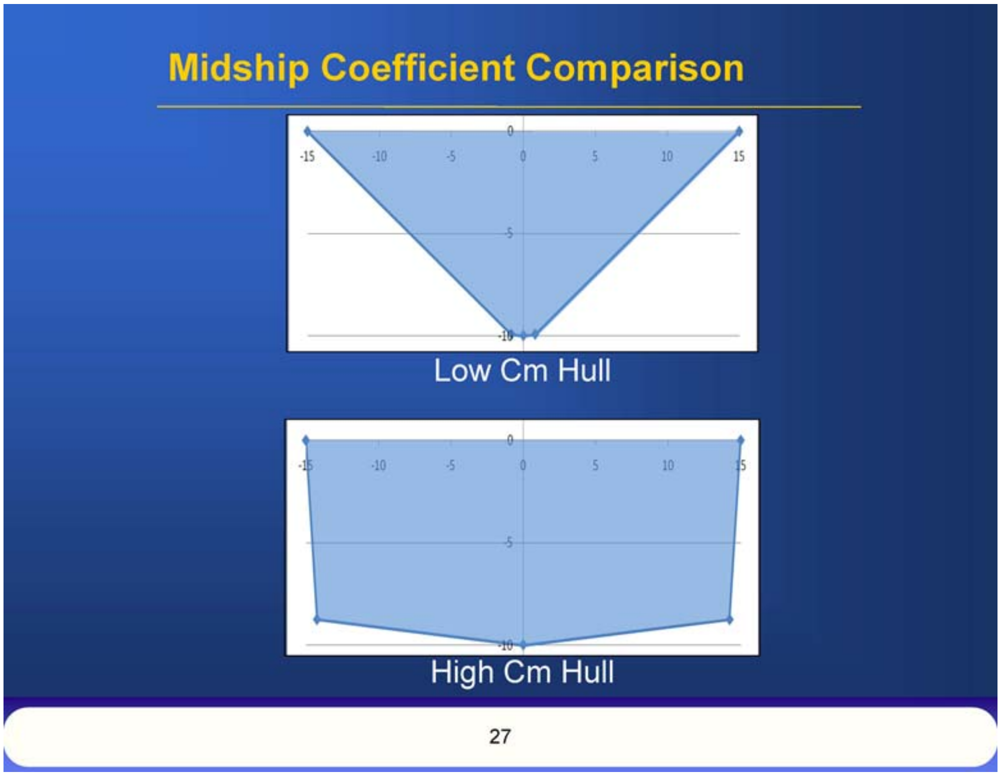
- Canadian Navy, URL: http://www.forces.gc.ca/site/_photos/orig/IS2009-6533.JPG [cited 29 July 2010].
- Jiang, C., Forstell, B., Lavis, D., and Ritter, O., "Ship hull and machinery optimization using physics-based design software," *Marine Technology and SNAME News*, Vol. 39, No. 2, Apr. 2002, pp. 109-117.
- Baum, S. J., and Ramakrishnan, R., "Applying 3D product modeling technology to shipbuilding," *Marine Technology and SNAME News*, Vol. 34, No. 1, Jan. 1997, pp. 56-65.
- Turner, M., "Lessons Learned from the GE90 3D Full Engine Simulations," *48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition*, AIAA 2010-1606-796, AIAA, Washington, DC, 2010, pp. 1-16.
- McKesson, C., "The Utility of Very Simple Models for Very Complex Systems," *2010 International Simulation Multiconference* [CD-ROM], Ottawa, Ontario, 12-14 Jul. 2010, pp. 181-187.
- Myers, R. H., Montgomery, D. C., Anderson-Cook, C. M., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 3rd ed., John Wiley & Sons, Inc., Hoboken, New Jersey, 2009, Chaps. 7, 8.

Appendix



DoE Runtime Environment

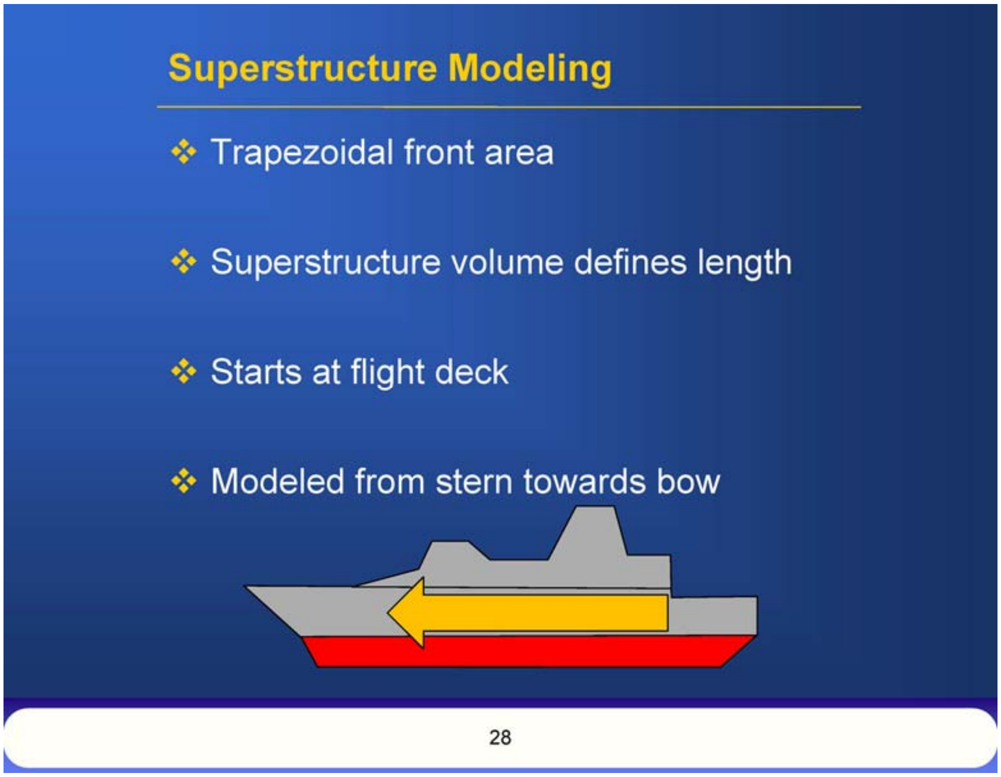
- Used to automatically test DoE cases



Midship Coefficient Comparison

Low Cm Hull

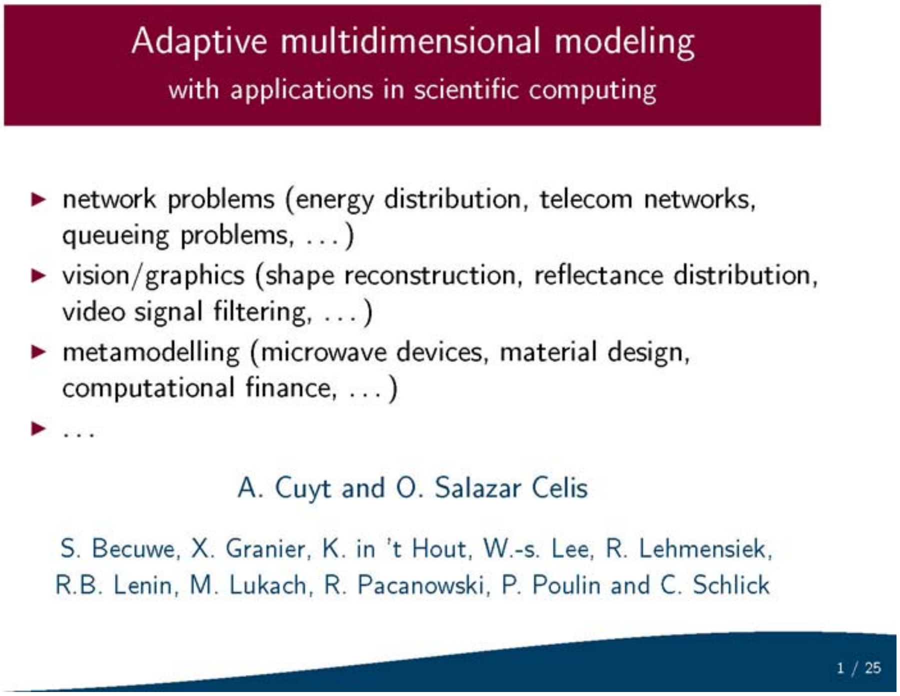
High Cm Hull



Superstructure Modeling

- Trapezoidal front area
- Superstructure volume defines length
- Starts at flight deck
- Modeled from stern towards bow

2.18 Adaptive multidimensional modeling with applications in scientific computing



Adaptive multidimensional modeling

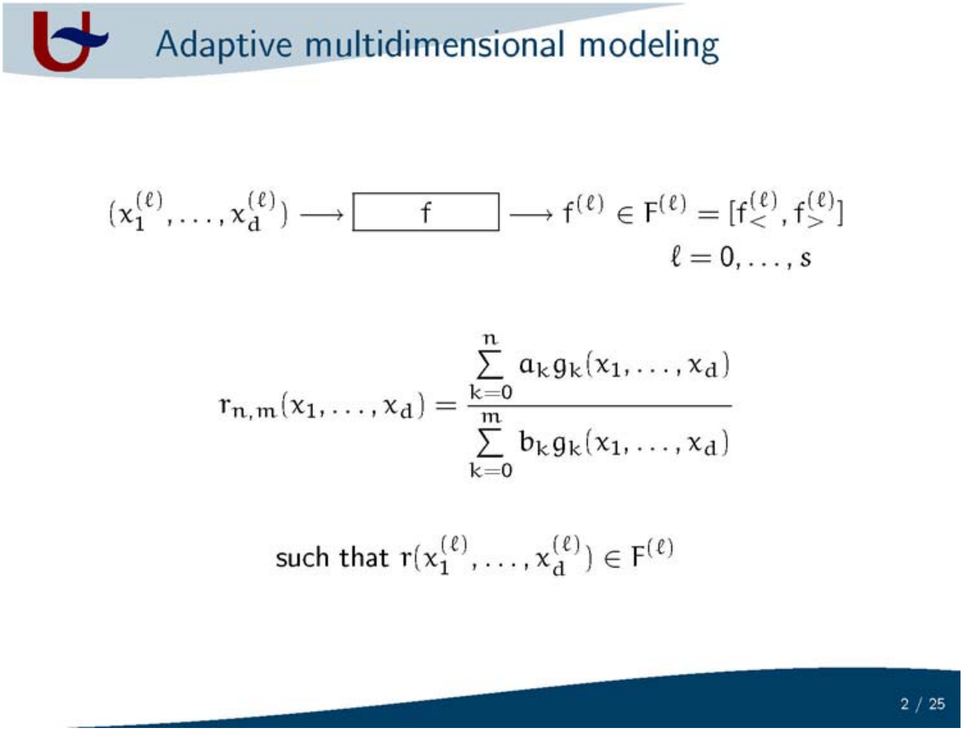
with applications in scientific computing

- network problems (energy distribution, telecom networks, queueing problems, ...)
- vision/graphics (shape reconstruction, reflectance distribution, video signal filtering, ...)
- metamodelling (microwave devices, material design, computational finance, ...)
- ...

A. Cuyt and O. Salazar Celis

S. Becuwe, X. Granier, K. in 't Hout, W.-s. Lee, R. Lehmensiek, R.B. Lenin, M. Lukach, R. Pacanowski, P. Poulin and C. Schlick

1 / 25



Adaptive multidimensional modeling

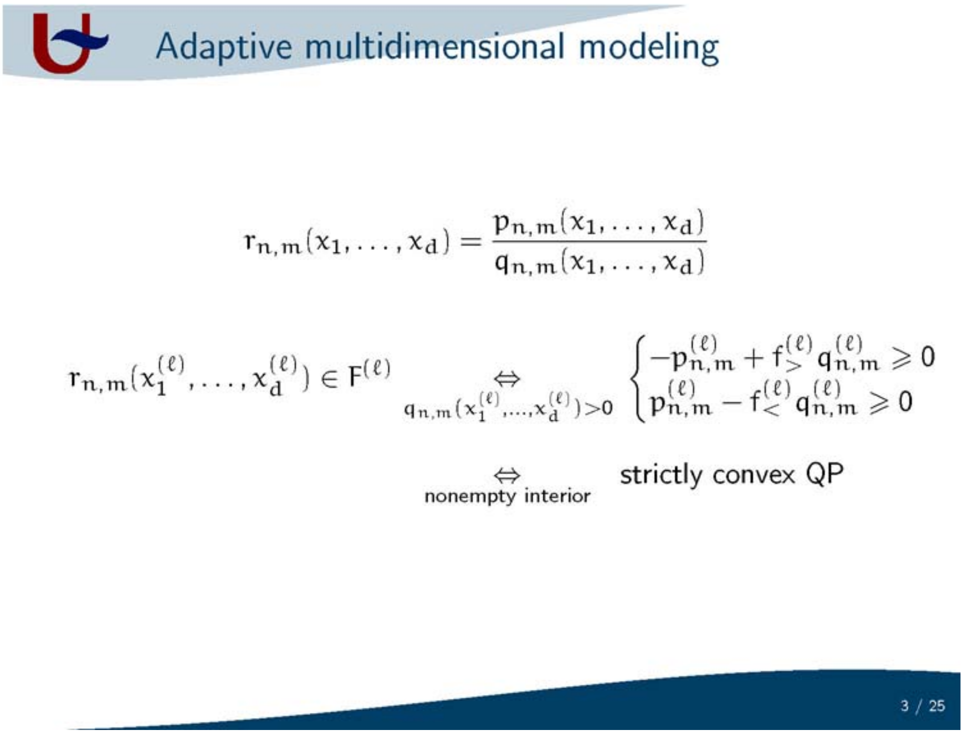
$$(x_1^{(\ell)},\dots,x_d^{(\ell)}) \longrightarrow \boxed{\quad f \quad} \longrightarrow f^{(\ell)} \in F^{(\ell)} = [f_<^{(\ell)}, f_>^{(\ell)}]$$

$$\ell = 0,\dots,s$$

$$r_{n,m}(x_1,\dots,x_d) = \frac{\sum_{k=0}^{n} a_k g_k(x_1,\dots,x_d)}{\sum_{k=0}^{m} b_k g_k(x_1,\dots,x_d)}$$

such that $r(x_1^{(\ell)},\dots,x_d^{(\ell)}) \in F^{(\ell)}$

2 / 25

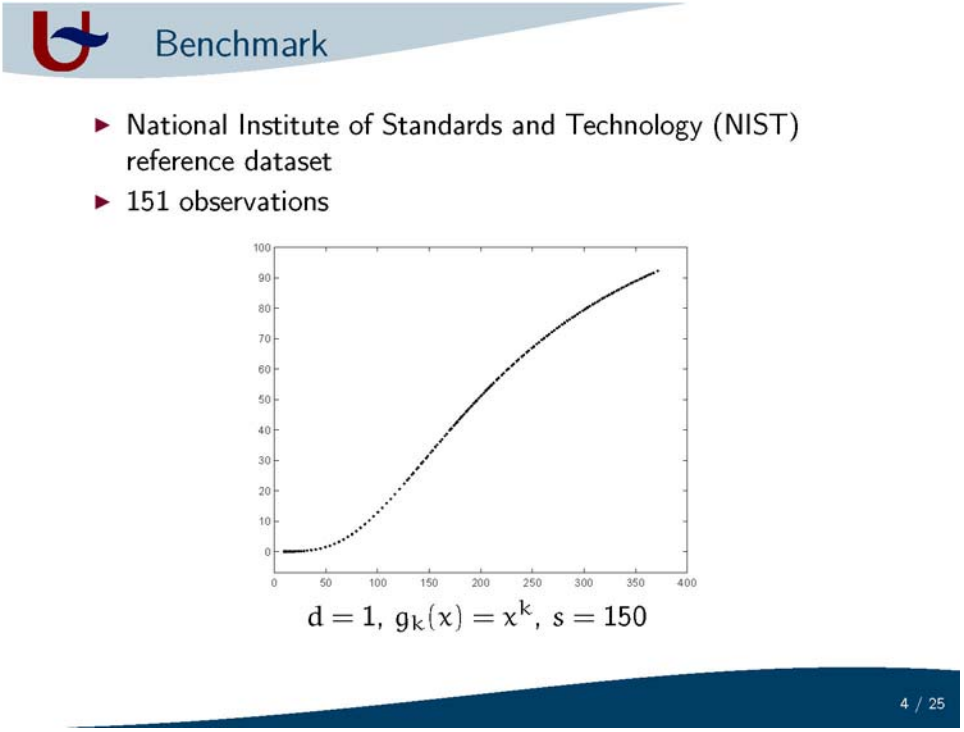


Adaptive multidimensional modeling

$$r_{n,m}(x_1,\dots,x_d)=\frac{p_{n,m}(x_1,\dots,x_d)}{q_{n,m}(x_1,\dots,x_d)}$$

$$r_{n,m}(x_1^{(\ell)},\dots,x_d^{(\ell)})\in F^{(\ell)} \underset{q_{n,m}(x_1^{(\ell)},\dots,x_d^{(\ell)})>0}{\Leftrightarrow} \begin{cases} -p^{(\ell)}_{n,m}+f^{(\ell)}_{>}q^{(\ell)}_{n,m}\geqslant 0\\ p^{(\ell)}_{n,m}-f^{(\ell)}_{<}q^{(\ell)}_{n,m}\geqslant 0\end{cases}$$

$$\underset{\text{nonempty interior}}{\Leftrightarrow} \quad \text{strictly convex QP}$$



Benchmark

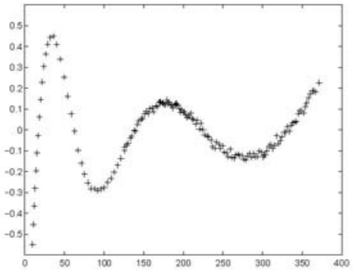
- National Institute of Standards and Technology (NIST) reference dataset
- 151 observations

$d=1,\ g_k(x)=x^k,\ s=150$

best ℓ_2-approximation

$$\sum_{\ell=0}^{s} \left(r^*_{2,2}(x_1^{(\ell)}, \ldots, x_d^{(\ell)}) - f^{(\ell)} \right)^2 \text{ minimal}$$

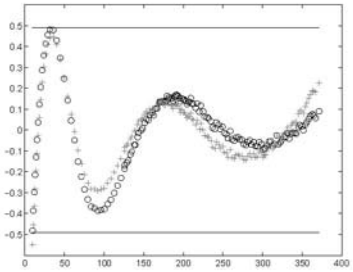
$$r^*_{2,2} = \frac{1.6745 - 0.13927x + 0.00260x^2}{1 - 0.00172x + 0.00002x^2}$$



residuals, $\sigma = 0.16355$

$$f_>^{(\ell)} - f_<^{(\ell)} = 2(3\sigma) = 0.9813 \Rightarrow n = 2, m = 2$$

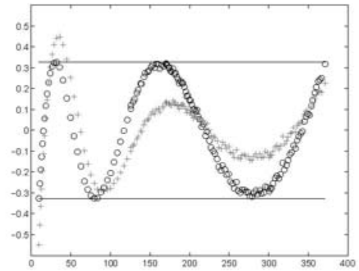
$$r_{2,2} = \frac{1.56271 - 0.13713x + 0.00261x^2}{1 - 0.00173x + 0.00002x^2}$$



residuals

$$f_{>}^{(\ell)} - f_{<}^{(\ell)} = 2(2\sigma) = 0.6542 \Rightarrow n = 2, m = 2$$

$$r_{2,2} = \frac{1.16217 - 0.1080x + 0.00224x^2}{1 - 0.00223x + 0.00002x^2}$$

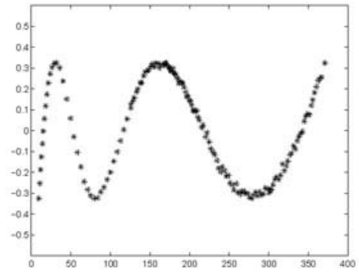


residuals

compare to best ℓ_∞-approximation

$$\max_{\ell=0,\ldots,s} \left| r_{2,2}^{\infty}(x_1^{(\ell)}, \ldots, x_d^{(\ell)}) - f^{(\ell)} \right| \text{ minimal}$$

$$r_{2,2}^{\infty} = \frac{1.15538 - 0.10751x + 0.00223x^2}{1 - 0.00223x + 0.00002x^2}$$

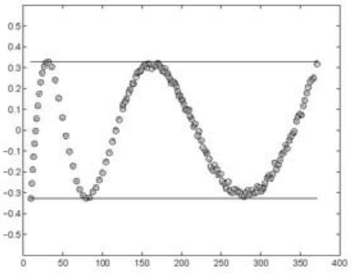


residuals, max = 0.3244

compare to best ℓ_∞-approximation

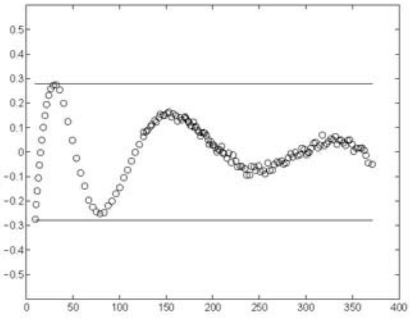
$$\max_{\ell=0,\dots,s}\left|r^{\infty}_{2,2}(x_1^{(\ell)},\dots,x_d^{(\ell)})-f^{(\ell)}\right| \text{ minimal}$$

$$r^{\infty}_{2,2}=\frac{1.15538-0.10751x+0.00223x^2}{1-0.00223x+0.00002x^2}$$



residuals

$$f^{(\ell)}_{>}-f^{(\ell)}_{<}=2(1.75\sigma)=0.5724\Rightarrow r^{*}_{2,2} \text{ and } r^{\infty}_{2,2} \text{ do not satisfy}$$
$$\Rightarrow n=3,\ m=2$$



residuals

Grid structured data

Ideal lowpass filter:

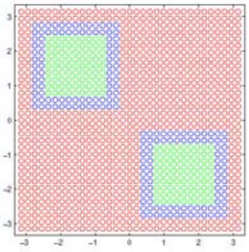
$$H\left(e^{it_1}, e^{it_2}\right) = \begin{cases} 1, & (t_1, t_2) \in P \subset [-\pi, \pi] \times [-\pi, \pi], \\ 0, & (t_1, t_2) \notin P. \end{cases}$$

In practice:

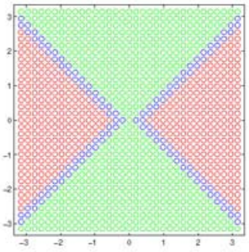
- passband $[1-\delta_1, 1+\delta_1]$, $(t_1, t_2) \in P$
- stopband $[-\delta_2, \delta_2]$, $(t_1, t_2) \notin P \cup T$
- transition band $[-\delta_2, 1+\delta_1]$, $(t_1, t_2) \in T$

Grid structured data

Examples of passband: $s + 1 = 33 \times 33$



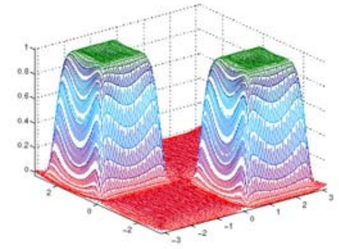
centro symmetric filter



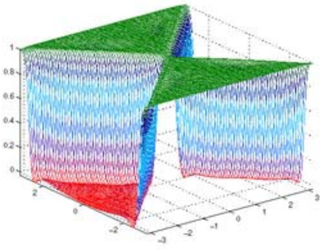
fan filter

Grid structured data

Rational models for the parameters $\delta_1 = 0.01$ and $\delta_2 = 0.02$



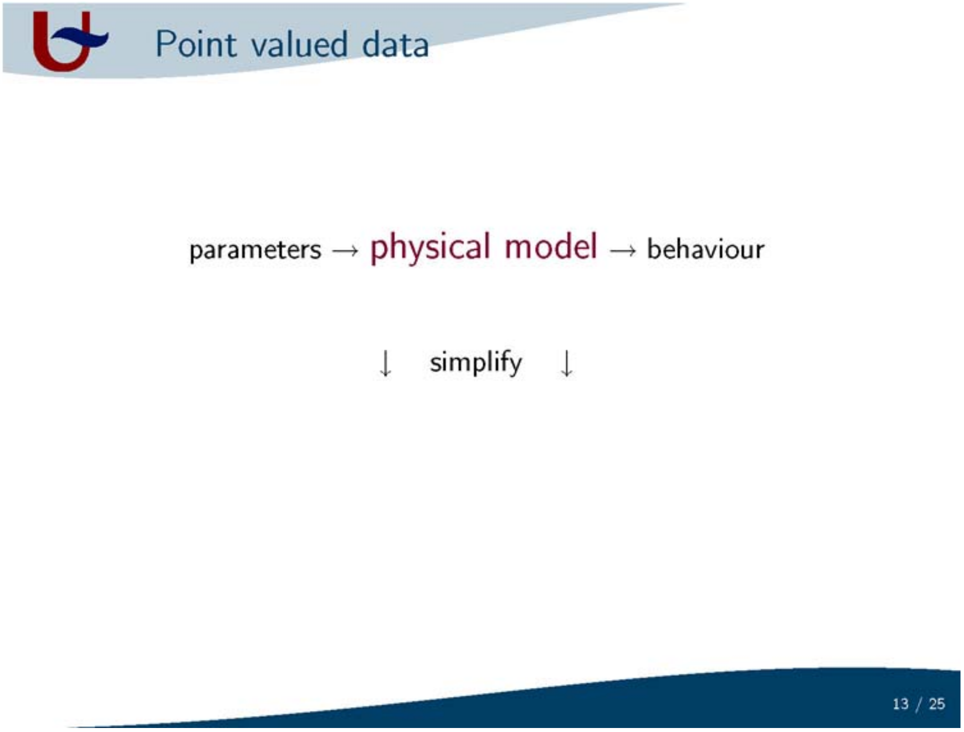
$r_{19,20}(t_1, t_2)$



$r_{12,14}(t_1, t_2)$

Point valued data

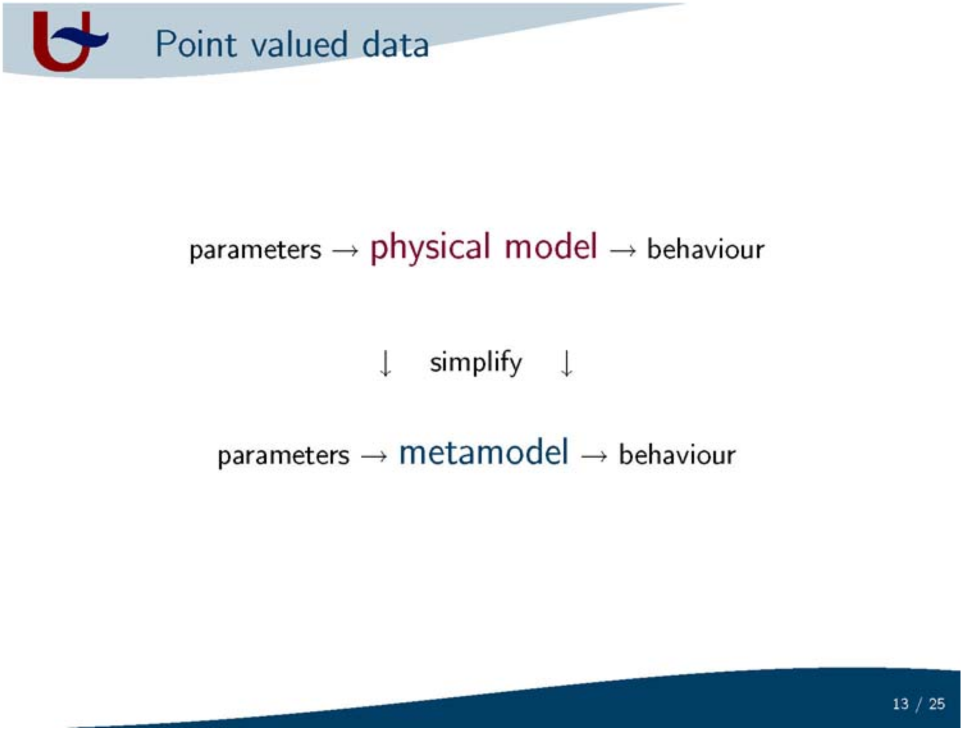
parameters → physical model → behaviour



Point valued data

parameters → physical model → behaviour

↓ simplify ↓

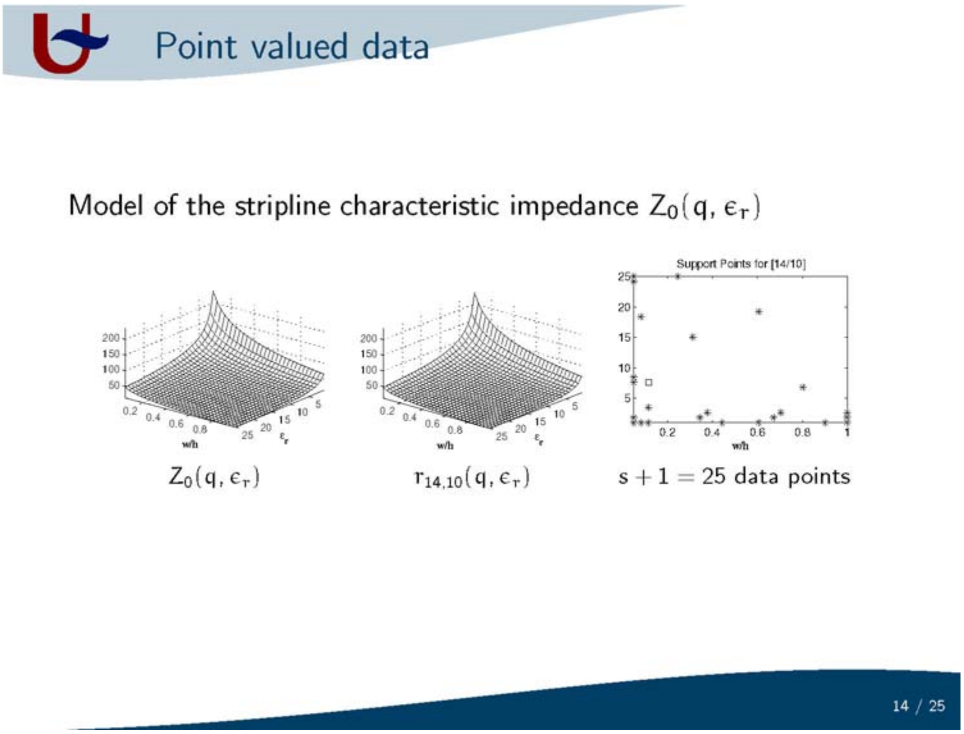


Point valued data

parameters → physical model → behaviour

↓ simplify ↓

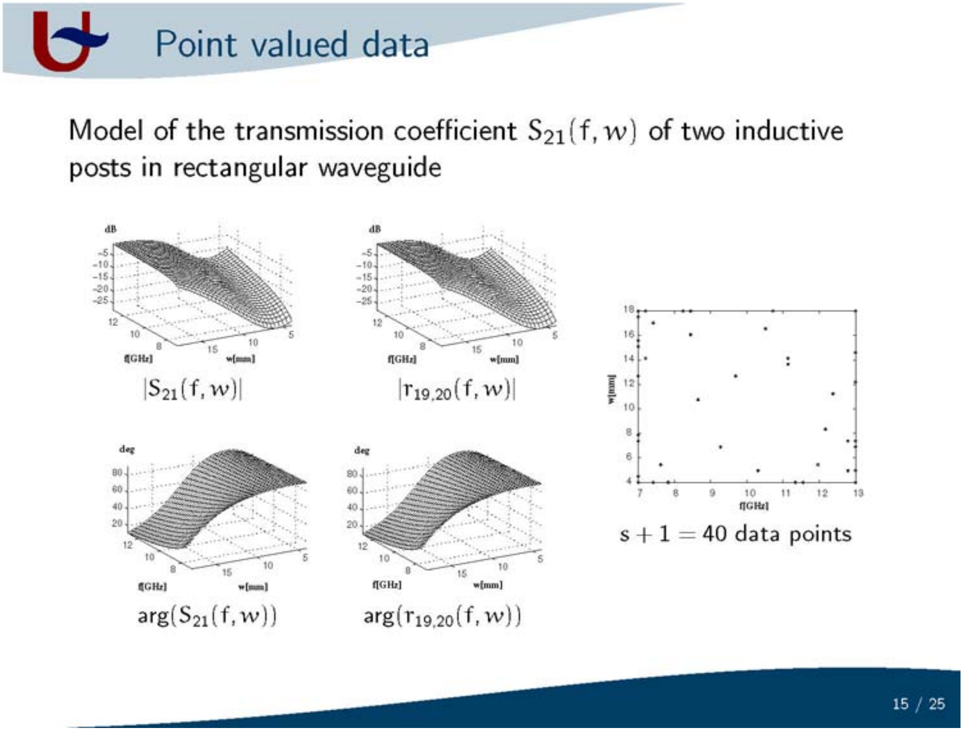
parameters → metamodel → behaviour



Point valued data

Model of the stripline characteristic impedance $Z_0(q, \epsilon_r)$

$Z_0(q, \epsilon_r)$ $r_{14,10}(q, \epsilon_r)$ $s + 1 = 25$ data points



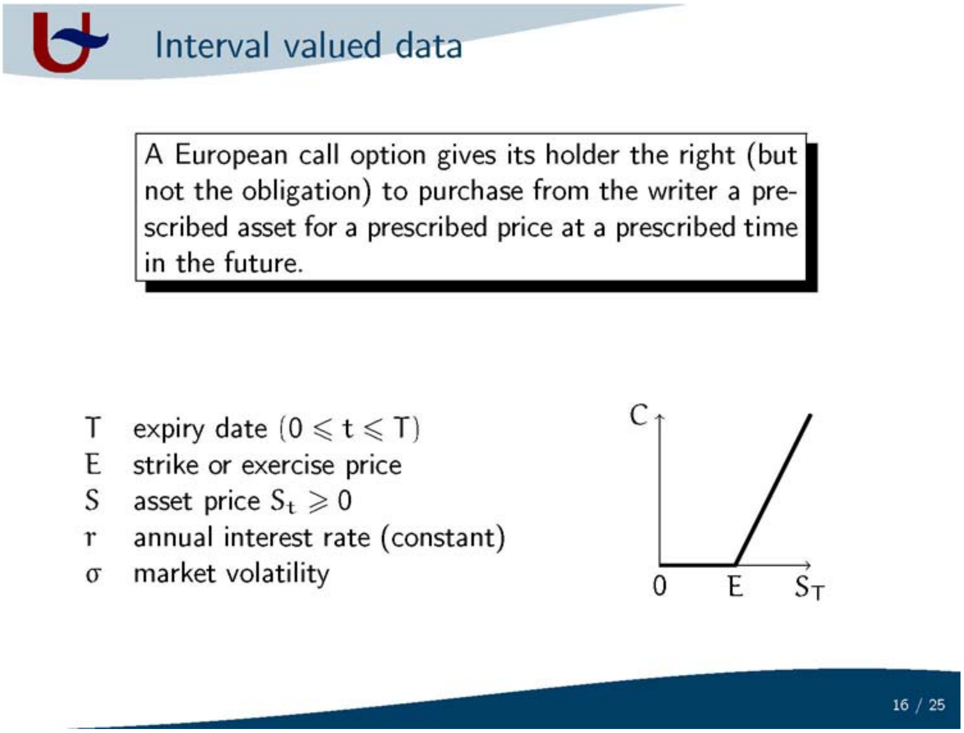
Point valued data

Model of the transmission coefficient $S_{21}(f, w)$ of two inductive posts in rectangular waveguide

$|S_{21}(f, w)|$ $|r_{19,20}(f, w)|$

$s + 1 = 40$ data points

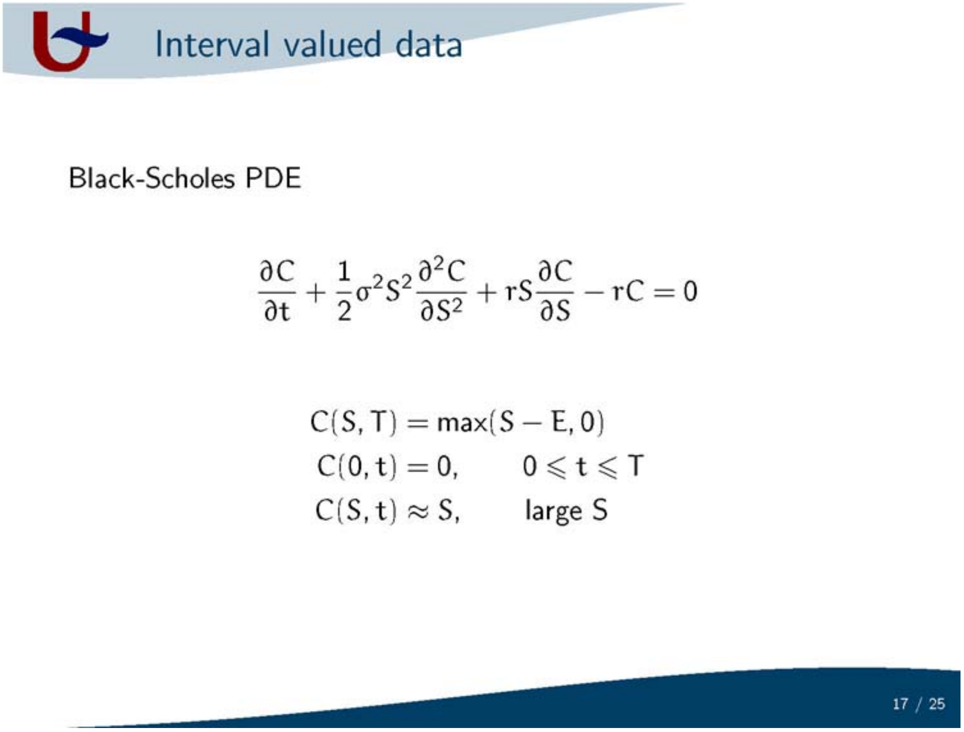
$\arg(S_{21}(f, w))$ $\arg(r_{19,20}(f, w))$



Interval valued data

A European call option gives its holder the right (but not the obligation) to purchase from the writer a prescribed asset for a prescribed price at a prescribed time in the future.

T	expiry date $(0 \leqslant t \leqslant T)$
E	strike or exercise price
S	asset price $S_t \geqslant 0$
r	annual interest rate (constant)
σ	market volatility

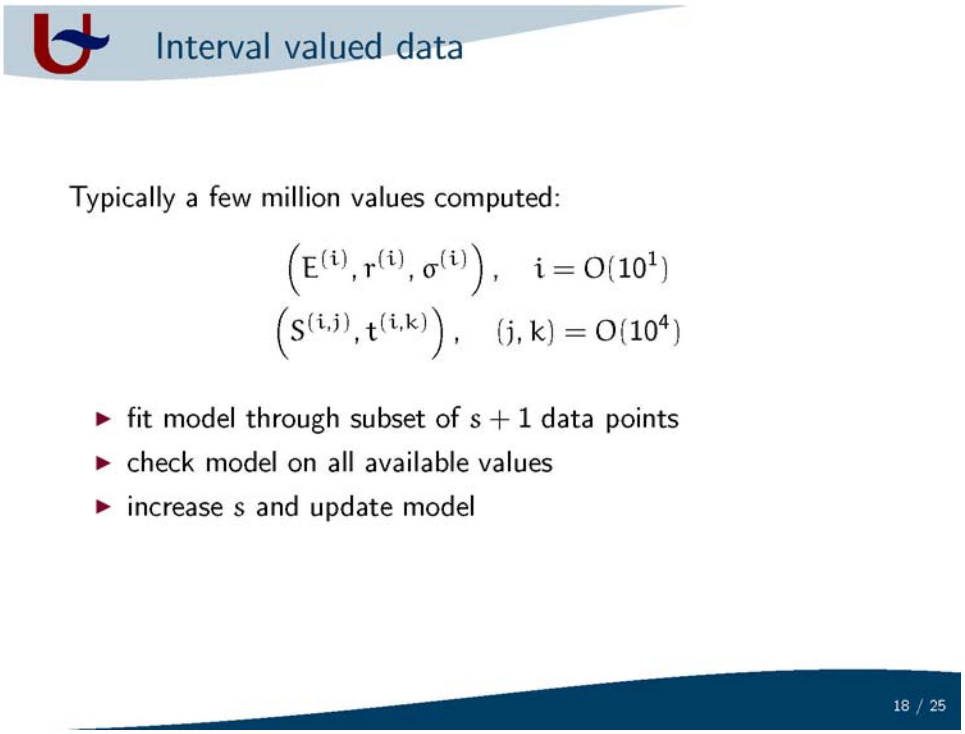


Interval valued data

Black-Scholes PDE

$$\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + rS\frac{\partial C}{\partial S} - rC = 0$$

$$\begin{aligned} C(S,T) &= \max(S-E,0) \\ C(0,t) &= 0, \qquad 0 \leqslant t \leqslant T \\ C(S,t) &\approx S, \qquad \text{large } S \end{aligned}$$



Interval valued data

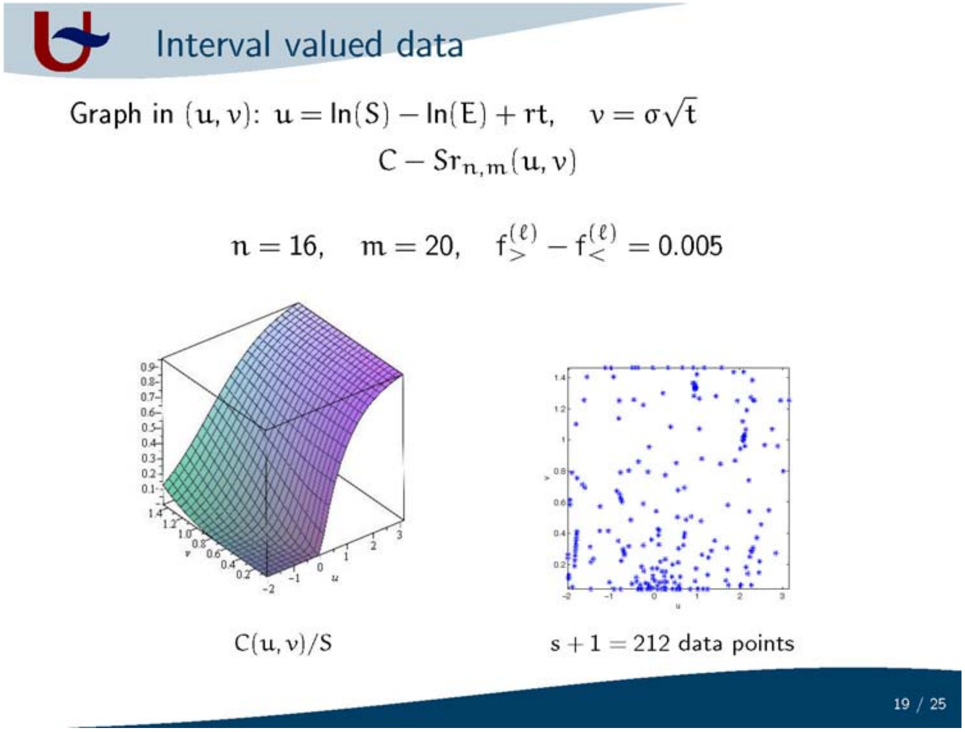
Typically a few million values computed:

$$\left(E^{(i)}, r^{(i)}, \sigma^{(i)}\right), \quad i = O(10^1)$$

$$\left(S^{(i,j)}, t^{(i,k)}\right), \quad (j,k) = O(10^4)$$

- fit model through subset of $s+1$ data points
- check model on all available values
- increase s and update model

18 / 25



Interval valued data

Graph in (u, v): $u = \ln(S) - \ln(E) + rt, \quad v = \sigma\sqrt{t}$

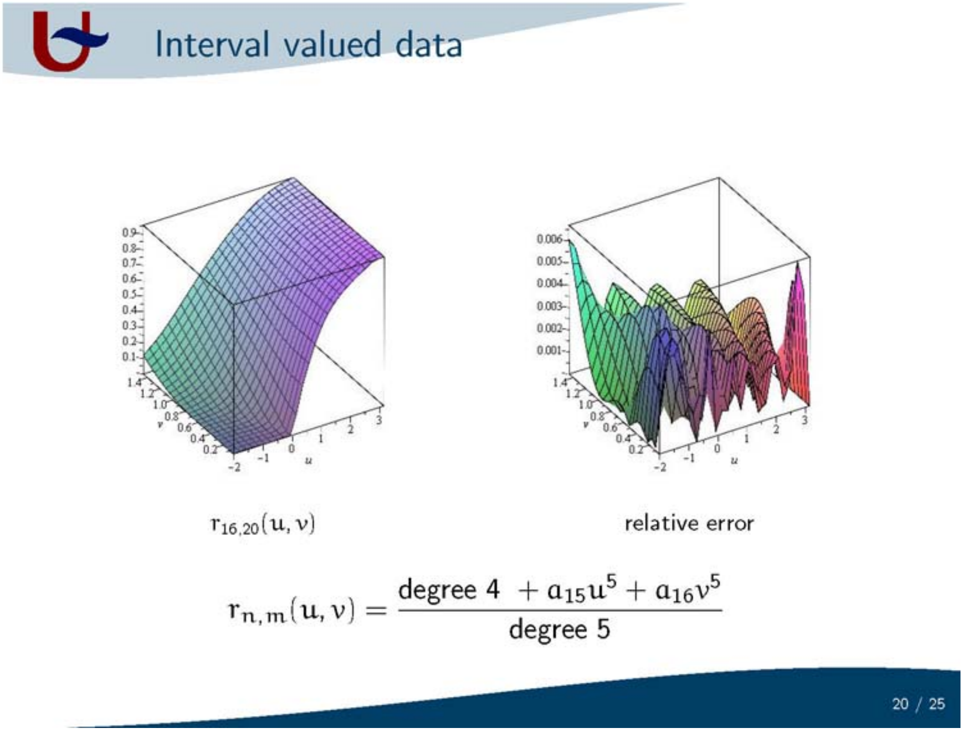
$$C - Sr_{n,m}(u, v)$$

$$n = 16, \quad m = 20, \quad f_{>}^{(\ell)} - f_{<}^{(\ell)} = 0.005$$

$C(u, v)/S$

$s + 1 = 212$ data points

19 / 25



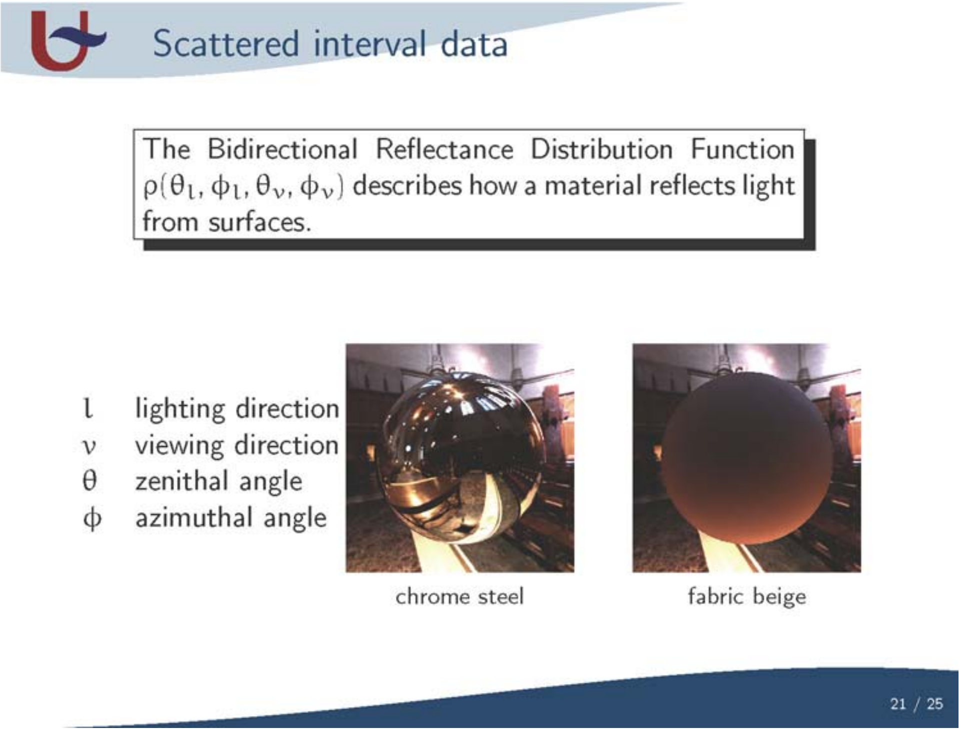
Interval valued data

$r_{16,20}(u,v)$

relative error

$$r_{n,m}(u,v) = \frac{\text{degree 4 } + a_{15}u^5 + a_{16}v^5}{\text{degree 5}}$$

20 / 25



Scattered interval data

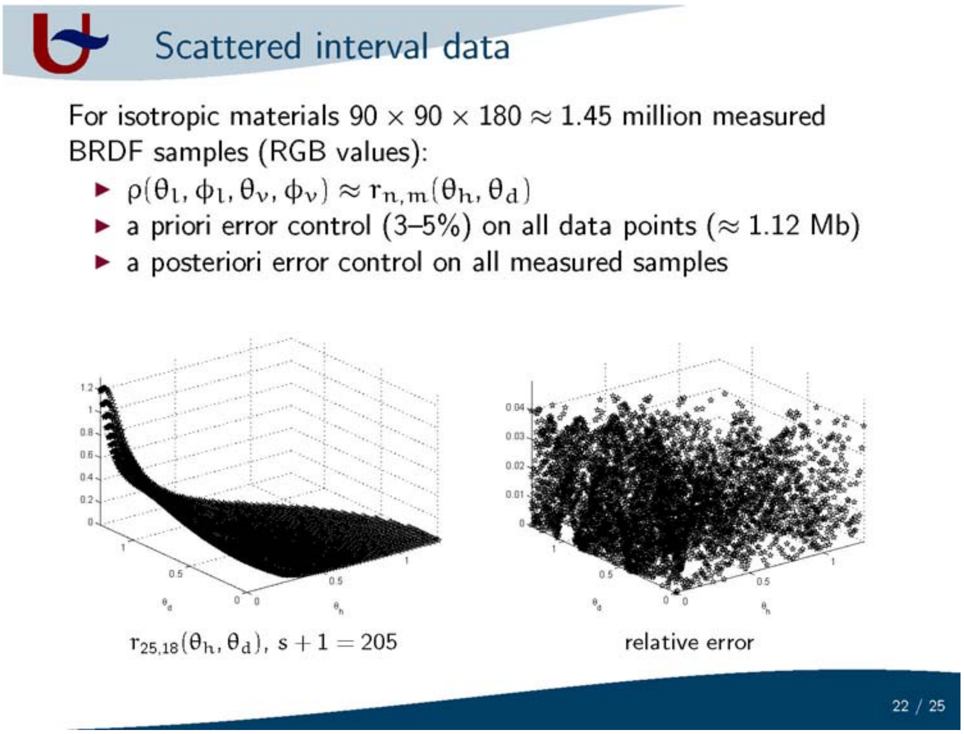
The Bidirectional Reflectance Distribution Function $\rho(\theta_l, \phi_l, \theta_v, \phi_v)$ describes how a material reflects light from surfaces.

l	lighting direction
v	viewing direction
θ	zenithal angle
ϕ	azimuthal angle

chrome steel

fabric beige

21 / 25



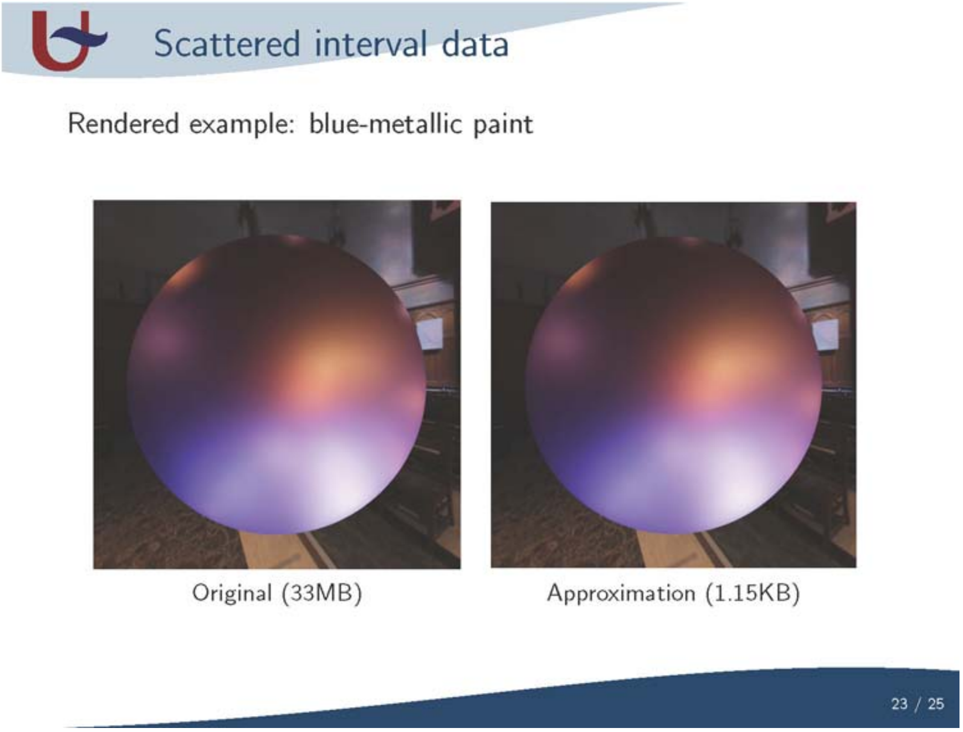
Scattered interval data

For isotropic materials 90 × 90 × 180 ≈ 1.45 million measured BRDF samples (RGB values):

- $\rho(\theta_l, \phi_l, \theta_v, \phi_v) \approx r_{n,m}(\theta_h, \theta_d)$
- a priori error control (3–5%) on all data points (≈ 1.12 Mb)
- a posteriori error control on all measured samples

$r_{25,18}(\theta_h, \theta_d)$, $s + 1 = 205$

relative error



Scattered interval data

Rendered example: blue-metallic paint

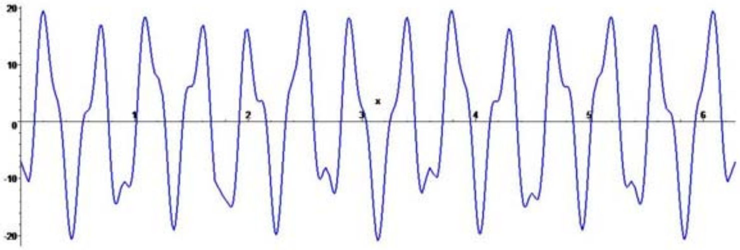
Original (33MB)

Approximation (1.15KB)

Sparsity

Compressive sensing recovers a K–sparse signal from only $M \approx K$ measurements without loss of information.

$$x(t) = 2\cos(5t) - 15\cos(14t) + \cos(26t) \qquad 0 \leqslant t \leqslant 2\pi$$
$$+ 5\cos(35t) + \text{noise}([-0.1, 0.1]),$$



Sparsity

$$t_j = j\frac{2\pi}{71}, \quad j = 0, \dots, 7, \qquad s + 1 = 8$$

Choice of datapoints allows to recover the frequencies first,

$$5 \quad 14 \quad 26 \quad 35$$

and afterwards the coefficients by fitting the same data,

$$2.004 \quad -14.93 \quad 0.9668 \quad 5.007$$

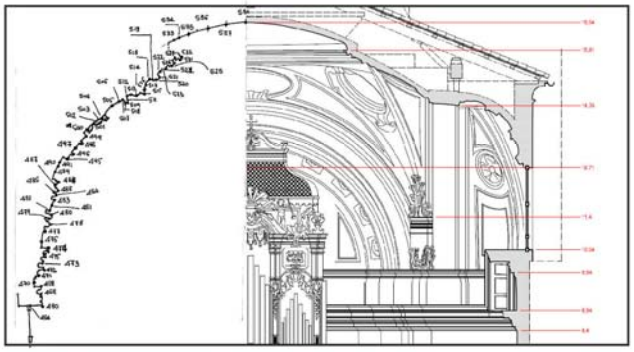
$$x(t) = 2\cos(5t) - 15\cos(14t) + \cos(26t) + 5\cos(35t)$$

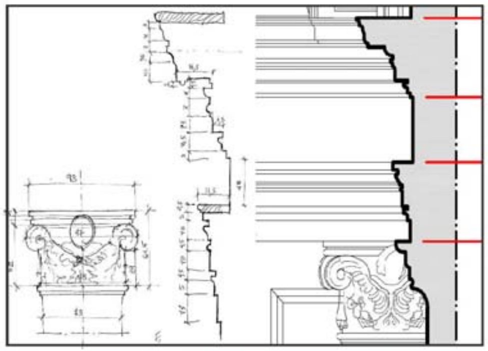
2.19 Three-dimensional model for preservation and restorationof architectural heritage

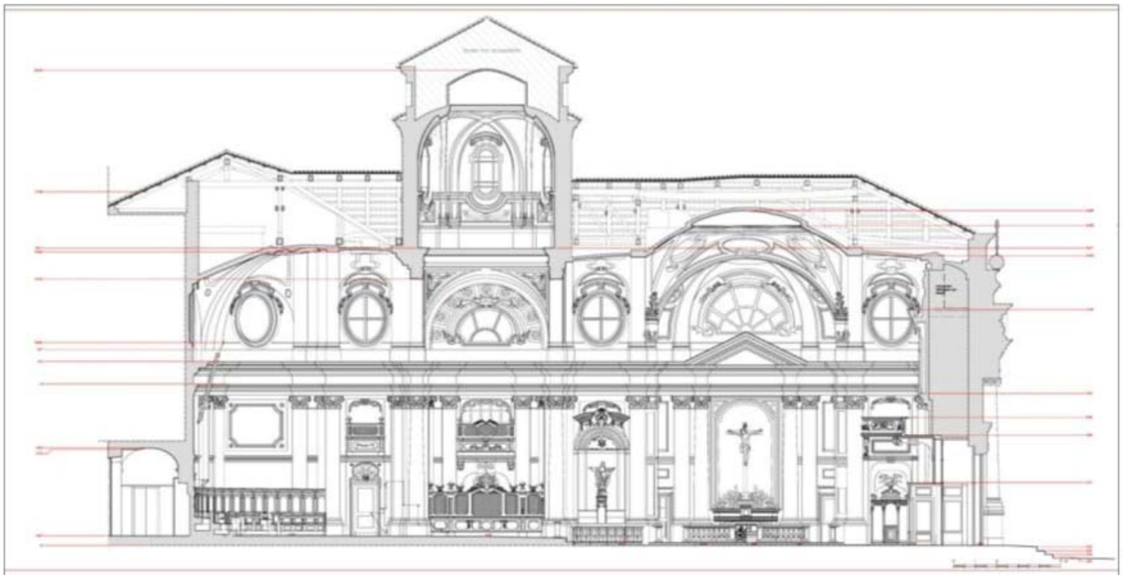
Three-dimensional model for preservation and restoration of architectural heritage

Elena Marchis,
PhD student, Politecnico di Torino, Departments of Building Engineering and Territorial Systems

The research focuses on developing guideline, creating a simple three-dimensional model designed to represent both the complexity of the "cultural heritage" morphology, as well as the need to manage the process of restoration in all its phases: from first findings to the restored final output.







The "Chiesa della Misericordia" case study

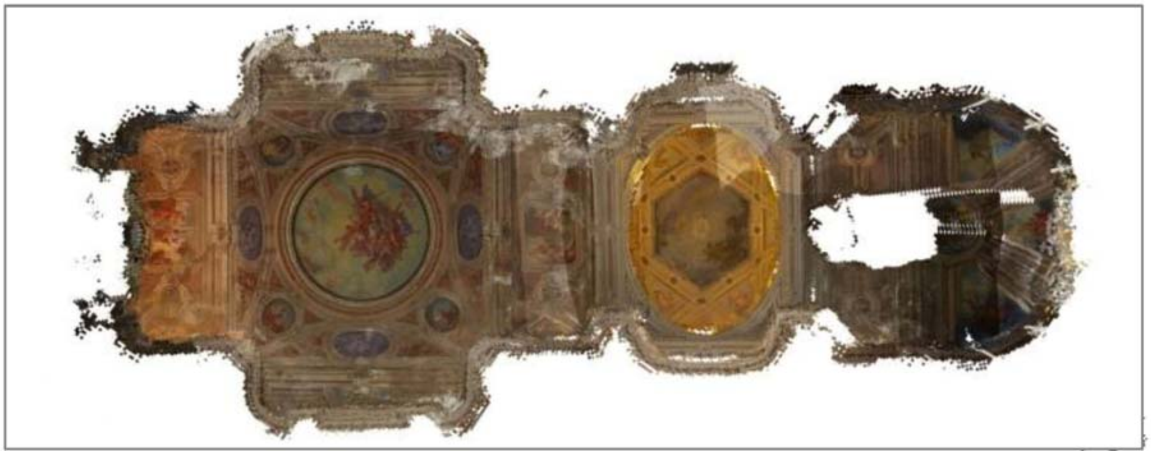
The first phase of the research was the architectural survey (in scale 1 to 50) and the graphical restitution of the baroque church of the Misericordia, located in Turin. The active participation to the survey was possible thanks to a collaborative project signed by the Department of Building Engineering and Territorial Systems of the Politecnico di Torino and the Confraternita della Misericordia, the Friary that owns the historical building. The research team directed by professor Secondino Coppo and composed by the engineers Bocconcino, Marchis, Piumatti, and Vitali was particularly careful to examine the following aspects:

- restitution of architectural details with direct metric techniques;
- interpretation of the cross section geometries starting from "total station" surveys, with eidotype support (Fig. B);
- profile tracing starting from the measured points "cloud", with a distance from the horizontal plane (accuracy of ± 2,5 cm);
- geometry reconstruction of the visible profiles starting from the solid image vectorialization and from the projection of the lines traced on the 3D model;
- redrawing of the architectonical elements starting from the digital images perspective correction;
- restitution of the building floor and vaulted ceiling plan, starting from the orthoimages. Some operating data of relieve.

The model has been completed notwithstanding the scarce archive documents and the practical difficulties of access to the upper parts of the hall. The detail level of the graphic information and the metric accuracy have been determined according to the content of the scale 1 to 50 drawing.

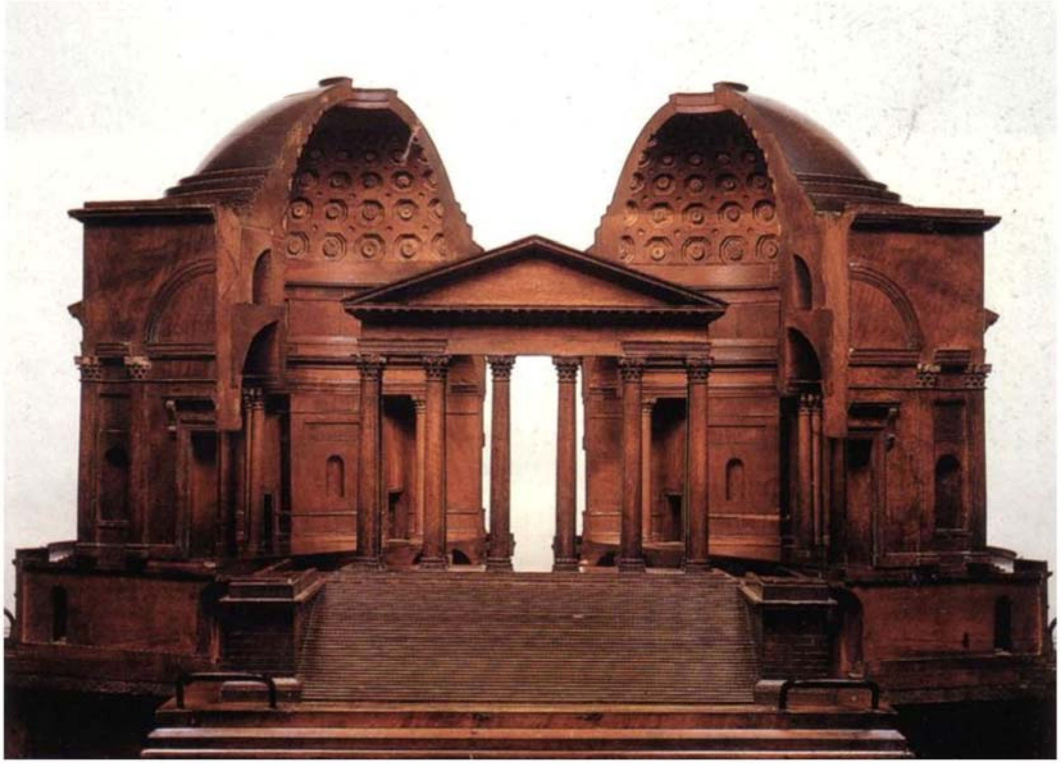
The operating times have been scheduled and followed according to the following scheme:

- 15 days for ground and first floor planes;
- 20 days for the longitudinal sections, in the lower part of the internal cornich);
- 15 days for the complete longitudinal sections;
- 7 days for the vault planes;
- 7 days for the details in the scale 1 to 20 (façade strip corresponding to the side fracture)



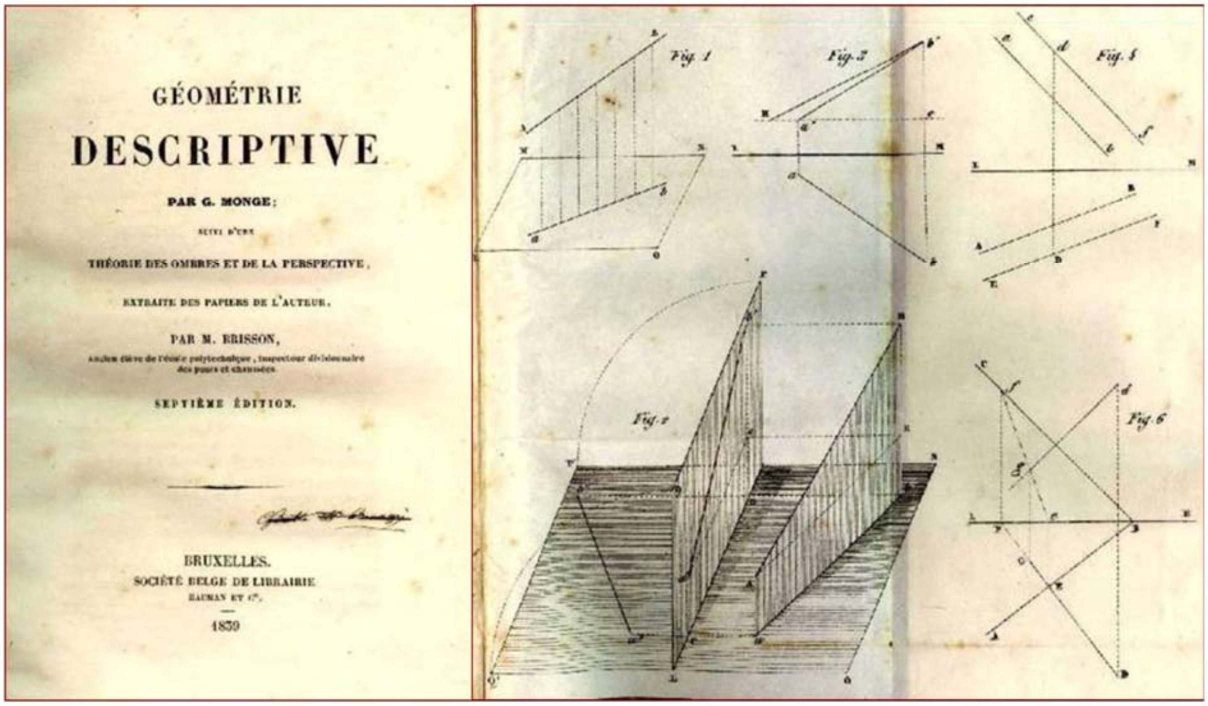
From sketching to the wooden models, and over.

Sketches, from the origin of the history are media, for transferring information and for centuries the drawing activities remained bounded to the figurative representation without giving the necessary data for an operational activity. In the mid Nineteenth Century Retdtenbacher , within the Industrial Revolution society, affirmed that technical drawing was the engineer can transfer his design thoughts in order to leave nothing to the free interpretation of the receiver of his message. But the drawing activity is important not only for designing but also for manufacturing the object with precision. The physical scale models, manufactured by architects and technicians are the only real mean to give concreteness to the design ideas.



Wooden model of the Church of the Gran Madre, Piedmontese woodcarver artisan according to a design of the architect Ferdinando Bonsignore, 1818, Torino, Museo Civico d'Arte Antica e Palazzo Madama (n. 1491/L)

Treatise illustrations of the Sixteenth and Seventeenth Centuries were pleasant pictures but were unable to bring the information for a definite practical realization of the object itself. Only with the birth of the *Descriptive Geometry* – an exact science bounded to the name of Gaspard Monge, a knowledge covered by military restrictions – the measure enters the quantitative graphical representation of the technical objects.



GÉOMÉTRIE

DESCRIPTIVE

PAR G. MONGE;

SUIVI D'UNE

THÉORIE DES OMBRES ET DE LA PERSPECTIVE,

EXTRAITE DES PAPIERS DE L'AUTEUR,

PAR M. BRISSON,

ancien élève de l'école polytechnique, inspecteur divisionnaire
des ponts et chaussées.

SEPTIÈME ÉDITION.

BRUXELLES.
SOCIÉTÉ BELGE DE LIBRAIRIE
HAUMAN ET Cᵉ.

1839

Gaspard Monge, (1839), *Géométrie descriptive*, Hauman, Bruxelles (VII ed.)

The Italian state of the art of a quantitative science.

A fundamental stage of the actual research performed was the evaluation of an effective need of a 3D tool for representing the complex morphology of the cultural heritage, but also for managing the complete restoration process during all its phases. Therefore an inquiry has been done in some architectural firms that operate in important restoration yards within the Piedmont region. This set of interviews was aimed to test and to verify the real and concrete possibilities of an application of the model in practical and professional activities, and for evaluating the cost/benefit frames. From these interviews emerged the model here presented, that has been directly applied to a real case study.

The search for structure in three phases:

- the first one is documentation and recording the state of the art

- the second one is focus on different applications and new methods of relief of the various stages of the building and mapping of degradation

- the third phase is to create a three dimensional model to be implemented with all the information gleaned in various stages of restoration, in order to return at the end of an entire overview of the evolution process of restoration, a three-dimensional database that contains information of different nature. The model could also be used to simulate more roads and fields for action, and then be used as decision support; be used to verify the stability and safety of buildings in relation to major structural movements and deformations

Stages of the building: application of the tear sheets for

One of the stages of the building of Misericordia was the degradation of the handcrafted mapping carried out by restorers. Specifically in the second phase is to propose the mapping of degradation not only on a stand-dimensional but a three-dimensional model. The need arises from the need to have detailed metric for quantifying the square meters of operations, work actually carried out or planned, and thus could not be offset by the actual surface projection representations, as in the presence of curved surfaces on time ....

The research aims to compare different methods and assess the cost-effectiveness.

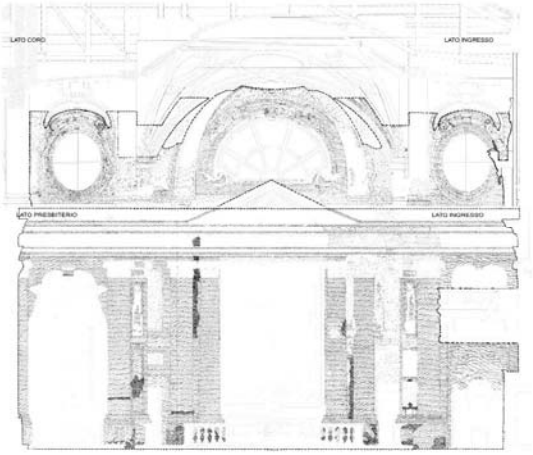
Methods compared:

a) traditional (hand mapping), two-dimensional (emphasis on very detailed)

b) digital two-dimensional (emphasis on very detailed)

c) digital three-dimensional (two-dimensional mapping model projecting three-dimensional)

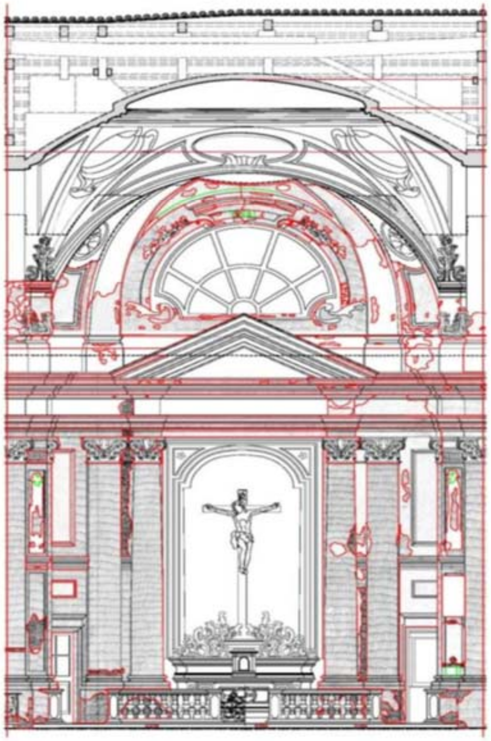
d) digital direct (directly on the picture solid), three-dimensional.



The proposal is to apply the traditional method applied to a three-dimensional model, starting from the classic procedure for mapping degradation, handmade on site by a restorer on a sound basis, as in the case of *Chiesa della Misericordia* in Turin (relieve scale 1:50).

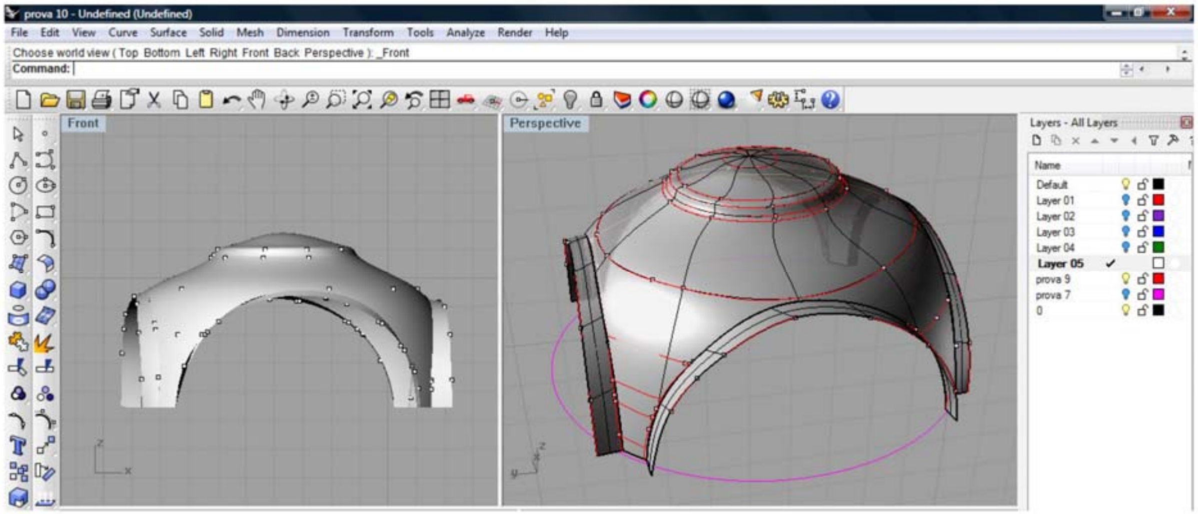
In the second stage it will be located throughout the map in digital form to facilitate the phases of reproduction, verification, calculation of the area.

The survey, in 1:50 scale, has created a three-dimensional model which will be projected on the regions, so that individual areas as close as possible to reality so as to also facilitate the calculation to quantify the areas affected by restoration.



The innovative part has been to propose the mapping of degradation not only on a stand-dimensional but a three-dimensional model, so you have no real offset surface representations in projection, as in the presence of curved surfaces, the times ....

Therefore created a three-dimensional model of the building is going to project areas mapped on the model. Need to be careful that the spread on three-dimensional surfaces coincide and that the level of errors are acceptable.

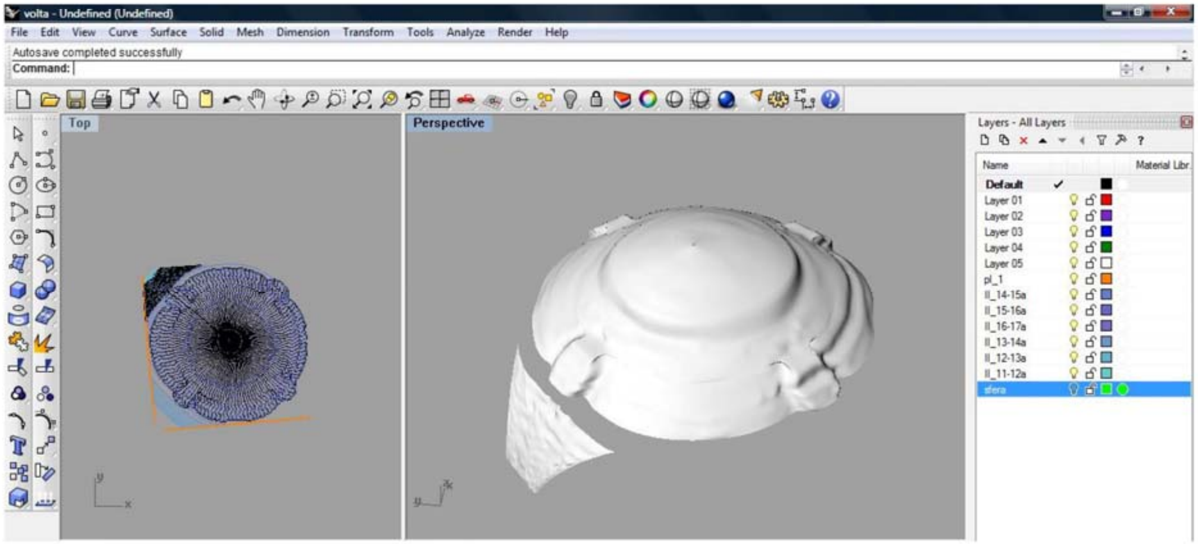


prova 10 - Undefined (Undefined)

```
Choose world view ( Top  Bottom  Left  Right  Front  Back  Perspective ): _Front
Command:
```

Layers - All Layers

Name	
Default	
Layer 01	
Layer 02	
Layer 03	
Layer 04	
Layer 05	✓
prova 9	
prova 7	
0	

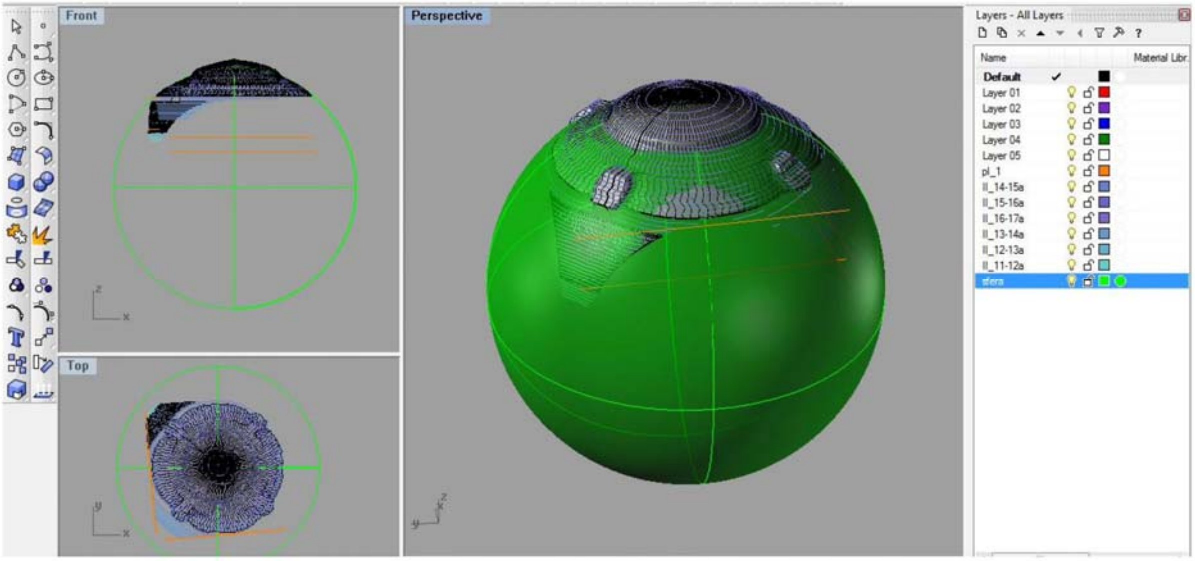


volta - Undefined (Undefined)

```
Autosave completed successfully
Command:
```

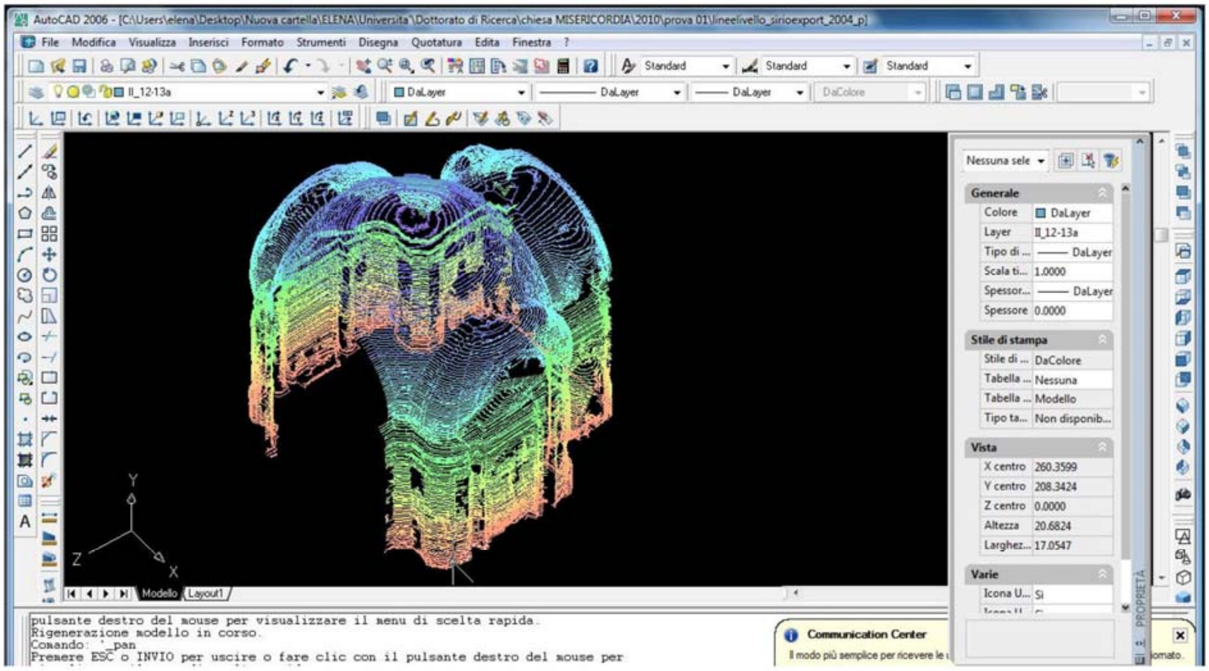
Layers - All Layers

Name		Material Libr.
Default	✓	
Layer 01		
Layer 02		
Layer 03		
Layer 04		
Layer 05		
pl_1		
ll_14-15a		
ll_15-16a		
ll_16-17a		
ll_13-14a		
ll_12-13a		
ll_11-12a		
sfera		



Layers - All Layers

Name		Material Libr.
Default	✓	
Layer 01		
Layer 02		
Layer 03		
Layer 04		
Layer 05		
pl_1		
ll_14-15a		
ll_15-16a		
ll_16-17a		
ll_13-14a		
ll_12-13a		
ll_11-12a		
sfera		


AutoCAD 2006 - [C:\Users\elena\Desktop\Nuova cartella\ELENA\Universita'\Dottorato di Ricerca\chiesa MISERICORDIA\2010\prova 01\lineelivello_sirioexport_2004_p]

Nessuna sele

Generale

Colore	DaLayer
Layer	II_12-13a
Tipo di ...	DaLayer
Scala ti...	1.0000
Spessor...	DaLayer
Spessore	0.0000

Stile di stampa

Stile di ...	DaColore
Tabella ...	Nessuna
Tabella ...	Modello
Tipo ta...	Non disponib...

Vista

X centro	260.3599
Y centro	208.3424
Z centro	0.0000
Altezza	20.6824
Larghez...	17.0547

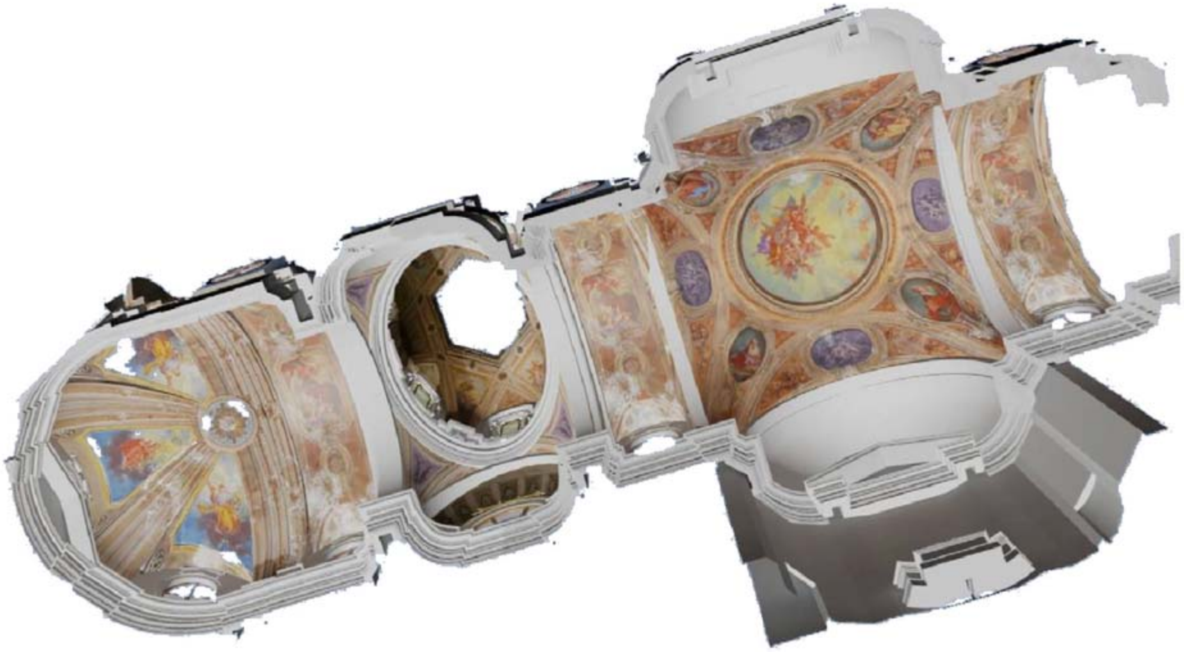
Varie

Icona U...	Sì

```
pulsante destro del mouse per visualizzare il menu di scelta rapida.
Rigenerazione modello in corso.
Comando: '_pan
Premere ESC o INVIO per uscire o fare clic con il pulsante destro del mouse per
```

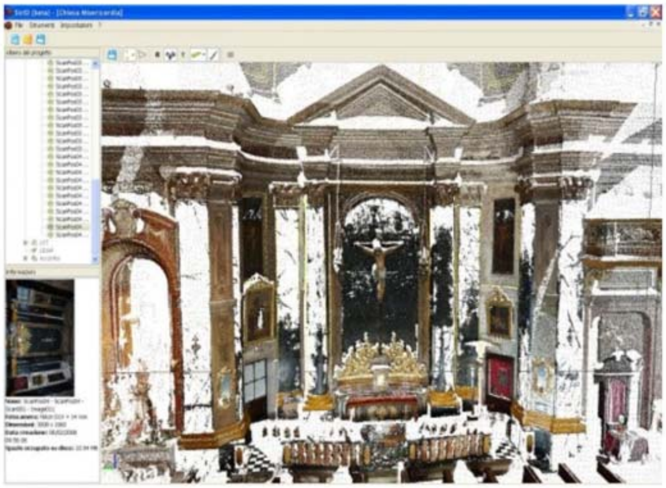
Communication Center

Il modo più semplice per ricevere le ... omato.

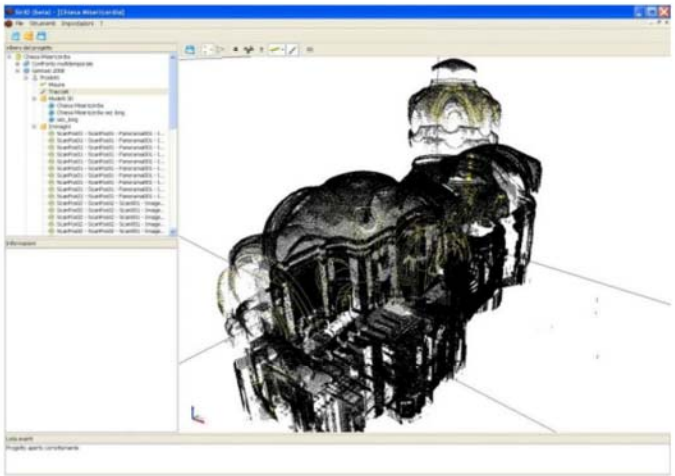


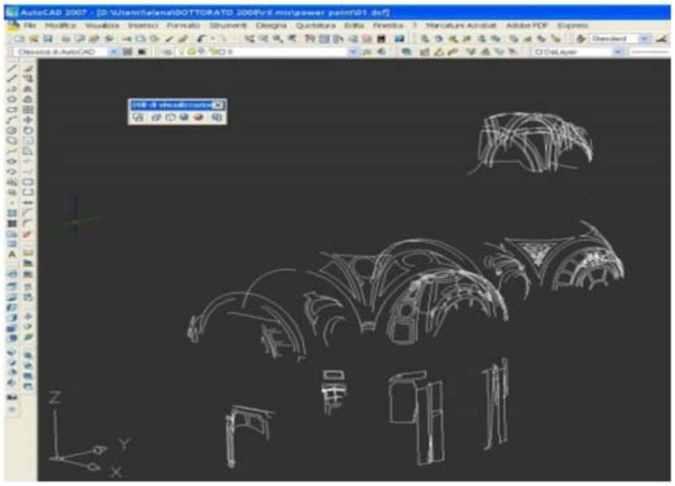
3D model based on architectural surveys in scale 1:50 performed by the team of Department of Building Engineering and Territorial Systems and Department of Land, Environment and Geo-Engineering.

The second method compares the mapping directly from a digital three-dimensional medium using the Image sound, with the help of the software Sirius. Polylines drawn solid image are exported and saved to each software in order to have the areas already mapped digitally and in three dimensions. This should reduce errors, speed up operations and cut costs.



The simple model can be constructed by integrating multiple technologies as demonstrated by laser scanners, photogrammetric survey, direct relief. The model also can collect the entire previous history of the property and subsequent transactions can happen in the future, the nature of maintenance work, information, etc.

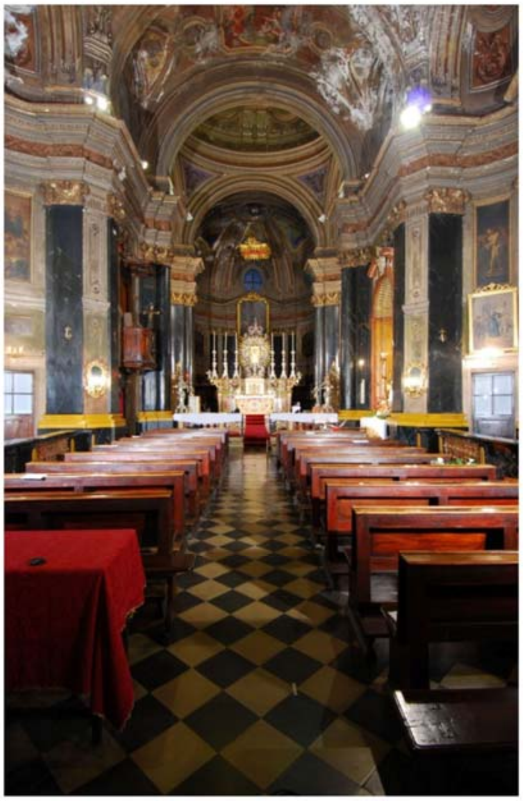


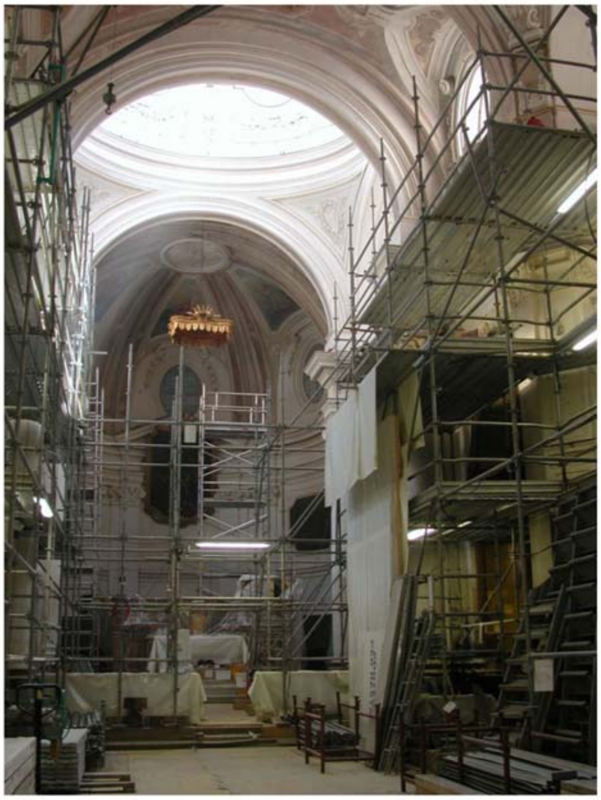

AutoCAD 2007
File Modifica Visualizza Inserisci Formato Strumenti Disegna Quotatura Edita Finestra ? Marcature Acrobat Adobe PDF Express
Classica di AutoCAD
Z
Y
X

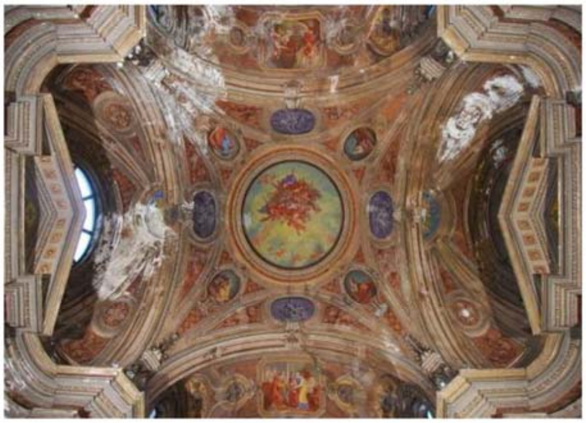
Test vector image sound - software testing Sirius (beta version) developed by SIR, Spin off of Politecnico di Torino.

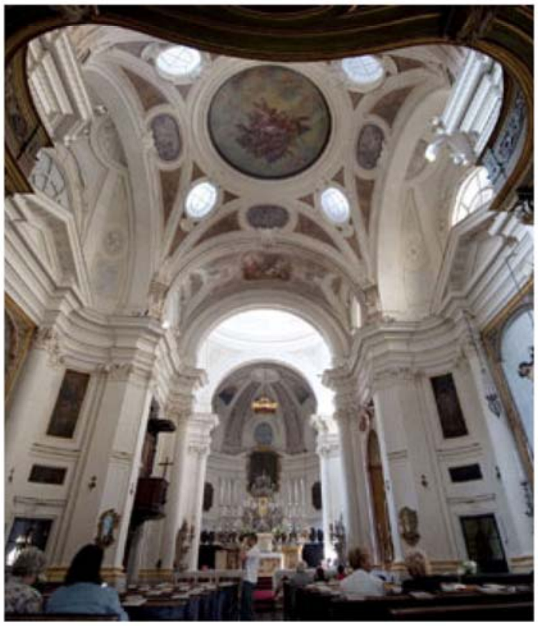
The aim of the research will be to create a model, three-dimensional mathematical, implementation, consultation and assistance to "large" restoration projects that will assist the structural analysis, allowing easier display of dynamic strain, analysis and lighting noise. It could also be a valuable tool for decision support, therefore, may simulate several possible scenarios for intervention. This model appears therefore an excellent support for recovering, ordering and monitoring information about materials and data (stage of restoration, photographs, sampling points, results of diagnostic tests, etc.) collected dynamically during the "life" of the cultural heritage, allowing to document its complete history.

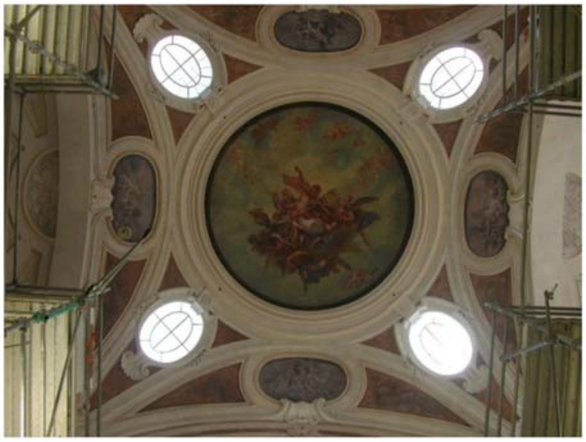
Several stages of construction: application of towels for ripping, tearing and opening of the "pentagon" after "ripped" the frescoes











Turin, Chiesa della Miseicordia, before and after restoration

References:

- Chiara Vernizzi, Considerazioni sul rilevamento per la valutazione strutturale: le volte della navata centrale del Duomo di Parma, in disegnare idee immagini n. 35/2007, pp.74-85.

- Beraldin e Marco Gaiani, Valutazioni delle prestazioni di sistemi di acquisizione tipo 3D active-vision: alcuni risultati, in DDD rivista trimestrale di Disegno digitale e design edita da Poli. Design. Anno 2, n. 5 Gen-mar 2003, pagg. 115-128.

- Lingua, A.; Piumatti P.; Rinaudo F., 2003. Digital photogrammetry: a standard approach to cultural eritage survey, in: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Ancona, Italy, Vol. XXXIV, Part 5/W12, pp. 210-215.

- Bornaz, L.; Dequal, S, 2003. The solid image: a new concept and its applications. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Ancona, Italy, Vol. XXXIV, Part 5/W12, pp. 78-82.

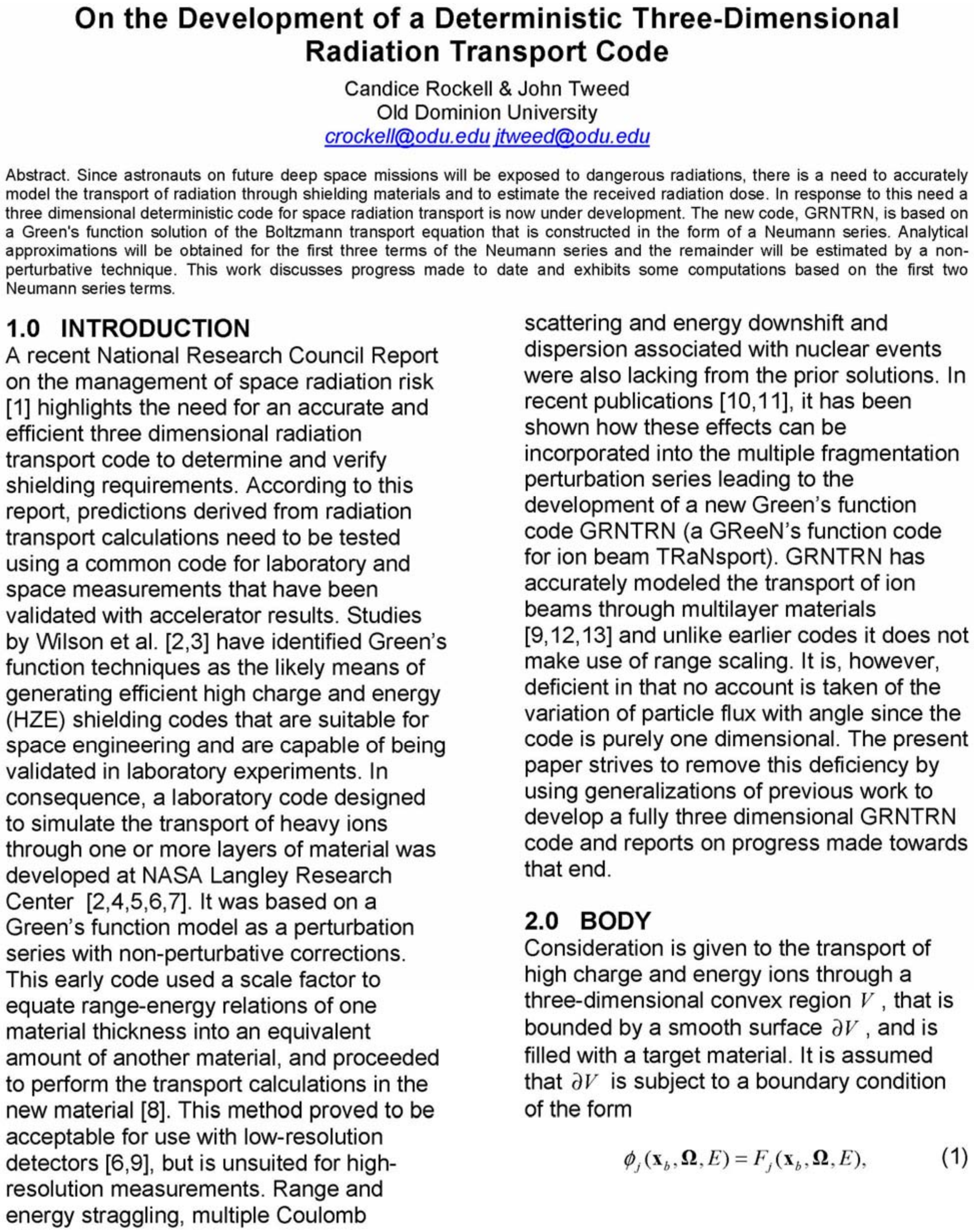
- Chiabrando, F.; Nex, F.; Piatti, D.; Rinaudo, F., 2008. Il rilievo metrico della Chiesa della Misericordia di Torino a supporto del cantiere di restauro. In: *Bollettino della Società Italiana di fotogrammetria e Topografia*, Vol. 1, pp 61-82.

- Marco Gaiani, Strategie di rappresentazione digitale: modelli per la conservazione e il restauro, in Quaderni Centro Ricerche Informatiche per i beni Culturali, Scuola Normale di Pisa, n. X, 2000, pagg. 47-69.

- Giorgio Veridiani, Il Battistero di San Giovanni in Piazza dei Miracoli a Pisa. Note sul rilievo digitale, in Firenze Architettura. Il disegno come conoscenza dell'architettura. Periodico semestrale del Dipartimento di progettazione dell'A rchitettura, Anno VII, n. 1& 2, pagg 32-39. Firenze, 2003.

- N. Milella, M. Zonno, A. Lerario, Il rilievo digitale dei beni architettonici., http://fabrica.ba.cnr.it/

2.20 On the Development of a Deterministic Three-Dimensional Radiation Transport Code



On the Development of a Deterministic Three-Dimensional Radiation Transport Code

Candice Rockell & John Tweed
Old Dominion University
crockell@odu.edu jtweed@odu.edu

Abstract. Since astronauts on future deep space missions will be exposed to dangerous radiations, there is a need to accurately model the transport of radiation through shielding materials and to estimate the received radiation dose. In response to this need a three dimensional deterministic code for space radiation transport is now under development. The new code, GRNTRN, is based on a Green's function solution of the Boltzmann transport equation that is constructed in the form of a Neumann series. Analytical approximations will be obtained for the first three terms of the Neumann series and the remainder will be estimated by a non-perturbative technique. This work discusses progress made to date and exhibits some computations based on the first two Neumann series terms.

1.0 INTRODUCTION

A recent National Research Council Report on the management of space radiation risk [1] highlights the need for an accurate and efficient three dimensional radiation transport code to determine and verify shielding requirements. According to this report, predictions derived from radiation transport calculations need to be tested using a common code for laboratory and space measurements that have been validated with accelerator results. Studies by Wilson et al. [2,3] have identified Green's function techniques as the likely means of generating efficient high charge and energy (HZE) shielding codes that are suitable for space engineering and are capable of being validated in laboratory experiments. In consequence, a laboratory code designed to simulate the transport of heavy ions through one or more layers of material was developed at NASA Langley Research Center [2,4,5,6,7]. It was based on a Green's function model as a perturbation series with non-perturbative corrections. This early code used a scale factor to equate range-energy relations of one material thickness into an equivalent amount of another material, and proceeded to perform the transport calculations in the new material [8]. This method proved to be acceptable for use with low-resolution detectors [6,9], but is unsuited for high-resolution measurements. Range and energy straggling, multiple Coulomb scattering and energy downshift and dispersion associated with nuclear events were also lacking from the prior solutions. In recent publications [10,11], it has been shown how these effects can be incorporated into the multiple fragmentation perturbation series leading to the development of a new Green's function code GRNTRN (a GReeN's function code for ion beam TRaNsport). GRNTRN has accurately modeled the transport of ion beams through multilayer materials [9,12,13] and unlike earlier codes it does not make use of range scaling. It is, however, deficient in that no account is taken of the variation of particle flux with angle since the code is purely one dimensional. The present paper strives to remove this deficiency by using generalizations of previous work to develop a fully three dimensional GRNTRN code and reports on progress made towards that end.

2.0 BODY

Consideration is given to the transport of high charge and energy ions through a three-dimensional convex region V, that is bounded by a smooth surface ∂V, and is filled with a target material. It is assumed that ∂V is subject to a boundary condition of the form

$$\phi_j(\mathbf{x}_b, \mathbf{\Omega}, E) = F_j(\mathbf{x}_b, \mathbf{\Omega}, E), \qquad (1)$$

where $F_j(\mathbf{x}_b,\mathbf{\Omega},E)$ is a prescribed function, and $\mathbf{x}_b$ is a point on the boundary. It is required that $\mathbf{\Omega}\cdot\mathbf{n}(x_b)<0$, where $\mathbf{n}(x_b)$ is the unit outward normal at $\mathbf{x}_b\in\partial V$, and the index *j* takes on values from 1 to N, where N is the number of ions in the model. $\phi_j(\mathbf{x},\mathbf{\Omega},E)$ is the flux of j-type ions with atomic mass A_j, at $\mathbf{x}\in V$, moving in the direction $\mathbf{\Omega}$ with energy E in units of MeV/amu.

2.1 Transport Theory

According to Wilson [14], the flux is given by the transport integral equation

$$\phi_j(\mathbf{x},\mathbf{\Omega},E)=\frac{P_j(\overline{E}_j)}{P_j(E)}\frac{\tilde{S}_j(\overline{E}_j)}{\tilde{S}_j(E)}F_j[\mathbf{x}'(\mathbf{x},\mathbf{\Omega}),\mathbf{\Omega},\overline{E}_j]$$

$$+\sum_{k>j}^{N}\int_{\rho'}^{\rho}\frac{P_j(E'')\tilde{S}_j(E'')d\rho''}{P_j(E)\tilde{S}_j(E)}\int_{E''}^{\infty}dE'\int_{4\pi}d\Omega'$$

$$\cdot\sigma_{jk}(\mathbf{\Omega},\mathbf{\Omega}',E'',E')\phi_k(\mathbf{x}_n+\rho''\mathbf{\Omega},\mathbf{\Omega}',E'),\quad(2)$$

where $\rho=\mathbf{x}\cdot\mathbf{\Omega}$, $\mathbf{x}'=\mathbf{x}-(\rho-\rho')\mathbf{\Omega}$ is the point where the ray through $\mathbf{x}$ in the direction $\mathbf{\Omega}$ enters V, $\tilde{S}_j(E)$ is the stopping power, $P_j(E)$ is the nuclear attenuation function and $\sigma_{jk}(\mathbf{\Omega},\mathbf{\Omega}',E,E')$ is the production cross section for j-type ions with energy E and direction $\mathbf{\Omega}$ by the collision of k-type ions with energy E' and direction $\mathbf{\Omega}'$. In addition $\overline{E}_j$ and E'' are defined by $\overline{E}_j=\overline{E}_j(\rho-\rho',E)\equiv R_j^{-1}[R_j(E)+\rho-\rho']$, and $E''=\overline{E}_j(\rho-\rho'',E)$ where $R_j(E)$ is the usual range-energy relation.

The production cross sections used in this paper are based on S. R. Blattnig's model, which is fully described in reference [17]. They are given by the approximation

$$\sigma_{jk}(\mathbf{\Omega},\mathbf{\Omega}_k,E,E_k)$$
$$\approx\sigma_{jk}(E_k)f_E(E,E_k)f_\Omega(\mathbf{\Omega},\mathbf{\Omega}_k,E_k),\quad(3)$$

where

$$f_E(E,E_k)=\frac{\exp\left(-\dfrac{(E_k-E_s-E)^2}{2\sigma^2\gamma_L{}^2\beta_L{}^2}\right)}{(2\pi)^{\frac{1}{2}}\sigma\gamma_L\beta_L},\quad(4)$$

$$f_\Omega(\mathbf{\Omega},\mathbf{\Omega}_k,E_k)=H[\pi/2-\theta]\frac{\gamma_L{}^2\beta_L{}^2m_p{}^2\cos\theta}{2\pi\sigma^2}$$

$$\cdot\exp\left(-\frac{\gamma_L{}^2\beta_L{}^2m_p{}^2\sin^2\theta}{2\sigma^2}\right).\quad(5)$$

and $\sigma_{jk}(E_k)$ is the total cross section. In these equations, m_p is the proton rest mass, γ_L and β_L are parameters associated with the Lorentz transformation from the fragment reference frame to the lab frame, $\theta=\cos^{-1}(\mathbf{\Omega}\cdot\mathbf{\Omega}_k)$ is the lab frame scattering angle, E_s is the lab frame energy shift, and σ the corresponding fragment momentum width [15,16].

By introducing the field vector

$$\mathbf{\Phi}(\mathbf{x},\mathbf{\Omega},E)=[\phi_j(\mathbf{x},\mathbf{\Omega},E)],\quad(6)$$

the fragmentation operator $\mathbf{\Xi}$

$$[\mathbf{\Xi}\cdot\mathbf{\Phi}]_j(\mathbf{x},\mathbf{\Omega},E)=\sum_{k>j}^{N}\int_{E}^{\infty}dE'\int_{4\pi}d\Omega'$$
$$\cdot\sigma_{jk}(\mathbf{\Omega},\mathbf{\Omega}',E,E')\phi_k(\mathbf{x},\mathbf{\Omega}',E'),\quad(7)$$

and the linear transport operator $\mathbf{L}$

$$[\mathbf{L}\cdot\mathbf{f}(\mathbf{\Omega}_1,E_1)]_j(\mathbf{x},\mathbf{x}'',\mathbf{\Omega},E)$$
$$=\frac{P_j(E'')}{P_j(E)}\frac{\tilde{S}_j(E'')}{\tilde{S}_j(E)}f_j(\mathbf{\Omega},E''),\quad(8)$$

the transport integral equation can be expressed in the operator form

$$\mathbf{\Phi}=\mathbf{G}^0\cdot\mathbf{F}+\mathbf{Q}\cdot\mathbf{L}\cdot\mathbf{\Xi}\cdot\mathbf{\Phi}\quad(9)$$

where $\mathbf{Q}$ represents the integral with respect to ρ''. Since Eq. (9) is a Volterra type integral equation, it admits the Neumann series solution [10]

$$\mathbf{\Phi} = \sum_{n=0}^{\infty} (\mathbf{Q} \cdot \mathbf{L} \cdot \mathbf{\Xi})^n \cdot \mathbf{G}^0 \cdot \mathbf{F} = \sum_{n=0}^{\infty} \mathbf{G}^n \cdot \mathbf{F} \qquad (10)$$

where $\mathbf{F}$ is the boundary flux vector and for $n \geq 1$,

$$\mathbf{G}^n = (\mathbf{Q} \cdot \mathbf{L} \cdot \mathbf{\Xi}) \cdot \mathbf{G}^{n-1}. \qquad (11)$$

In this solution, the term $\mathbf{G}^0 \cdot \mathbf{F}$ represents the primary flux vector and the term $\mathbf{G}^n \cdot \mathbf{F}$ represents the flux of the n^{th} generation of secondary ions produced.

When the boundary condition (1) takes the special form

$$\phi_j(\mathbf{x}_b, \mathbf{\Omega}, E) = \frac{\delta_{jm}}{2\pi} \delta(1 - \mathbf{\Omega} \cdot \mathbf{\Omega}_0) \cdot \delta(E - E_0) \overline{\delta}(\mathbf{x}_b - \mathbf{x}_0) \qquad (12)$$

where $\overline{\delta}$ is the 'surface delta function' on ∂V, the solution of Eq. (2) is called the Green's function and is denoted by the symbol $G_{jm}(\mathbf{x}, \mathbf{x}_0, \mathbf{\Omega}, \mathbf{\Omega}_0, E, E_0)$. Once the Green's function is known, the solution for an arbitrary boundary condition (1) can be obtained from the formula

$$\phi_j(\mathbf{x}, \mathbf{\Omega}, E) = [\mathbf{G} \cdot \mathbf{F}]_j(\mathbf{x}, \mathbf{\Omega}, E) = \sum_{k \geq j}^{N} \int_{\partial V} d\mathbf{x}_0 \int_{4\pi} d\Omega_0 \int_E^{\infty} dE_0 \cdot G_{jk}(\mathbf{x}, \mathbf{x}_0, \mathbf{\Omega}, \mathbf{\Omega}_0, E, E_0) F_k(\mathbf{x}_0, \mathbf{\Omega}_0, E_0). \qquad (13)$$

The summation is taken over $k \geq j$ (instead of $k > j$) to account for the primary ion spectrum.

2.2 The Zero Order Green's Function

The zero order Green's function is the first term in the Neumann series (10) with the unit boundary condition (12). On taking account of energy straggling, as described in [10] and [11], it can be shown that the zero order Green's function takes the form

$$G_{jm}^0(\mathbf{x}, \mathbf{x}_0, \mathbf{\Omega}, \mathbf{\Omega}_0, E, E_0) = \frac{P_m(\overline{E}_m)}{P_m(E)} \frac{\delta_{jm}}{2\pi} \frac{\delta(1 - \mathbf{\Omega} \cdot \mathbf{\Omega}_0) \overline{\delta}(\mathbf{x}' - \mathbf{x}_0)}{\sqrt{2\pi}\, s_m(\rho - \rho', E_0)} \cdot \exp\left\{ -\frac{[E - \hat{E}_m(\rho - \rho', E_0)]^2}{2 s_m(\rho - \rho', E_0)^2} \right\} \qquad (14)$$

where $\overline{E}_m = \overline{E}_m(\rho - \rho', E)$ and, by definition, $\hat{E}_m = \hat{E}_m(\rho - \rho', E_0) \equiv R_m^{-1}[R_m(E_0) - (\rho - \rho')]$ is the mean energy at depth $(\rho - \rho')$ g/cm^2 of an m-type ion that entered the transport material with energy E_0 MeV/amu, and $s_m(\rho - \rho', E_0)$ is the corresponding energy straggling width.

When the boundary condition takes the more general form (1), the primary flux is given by

$$\phi_j^0(\mathbf{x}, \mathbf{\Omega}, E) = [\mathbf{G}^0 \cdot \mathbf{F}]_j(\mathbf{x}, \mathbf{\Omega}, E) = \int_{\partial V} d\mathbf{x}_0 \int_{4\pi} d\Omega_0 \int_E^{\infty} dE_0 \cdot G_{jj}^0(\mathbf{x}, \mathbf{x}_0, \mathbf{\Omega}, \mathbf{\Omega}_0, E, E_0) F_j(\mathbf{x}_0, \mathbf{\Omega}_0, E_0), \qquad (15)$$

which, in general, needs to be evaluated numerically. In the accelerator beam model described below however, Eq. (15) can be approximated analytically and a closed form expression obtained for the zero order primary flux. The result obtained in this case is called the broad zero order Green's function.

Since ion beam experiments play an important role in analyzing the shielding requirements against dangerous space radiations, there is interest in modeling the propagation of linear accelerator beams through potential shielding materials. A simple model can be constructed by assuming that the accelerator beam consists of m-type ions with mean energy E_0 MeV/amu and mean direction $\mathbf{\Omega_0}$. It is further assumed that the beam has

Gaussian profiles in both angle and energy and that it enters the material at points that are distributed in a Gaussian manner about the mean point of entry $\mathbf{x}_0$. In order to accomplish this, it is assumed that the boundary ∂V is defined by the single-valued, continuously differentiable parametric equations $\partial V = \{\mathbf{x} : \mathbf{x} = \mathbf{x}(u,v), u_s \le u \le u_f, v_s \le v \le v_f\}$, in which case the element of surface area is given by $dS = |\partial_u \mathbf{x} \times \partial_v \mathbf{x}| du dv$ and the surface delta function is given by $\bar{\delta}(\mathbf{x} - \mathbf{x}_0) = |\partial_u \mathbf{x}_0 \times \partial_v \mathbf{x}_0|^{-1} \delta(u - u_0)\delta(v - v_0)$. The boundary condition (1) may then be assumed to take the Gaussian form

$$F_j(\mathbf{x}_b, \mathbf{\Omega}, E) = \frac{\delta_{jm}}{4\pi s_x{}^2 s_\Omega s_E K_x K_\Omega |\partial_u \mathbf{x}_b \times \partial_v \mathbf{x}_b|}$$

$$\cdot \exp\left\{-\frac{(u_b - u_0)^2 + (v_b - v_0)^2}{2 s_x{}^2}\right\} \exp\left\{-\frac{(E - E_0)^2}{2 s_E{}^2}\right\}$$

$$\cdot \exp\left\{-\frac{(1 - \mathbf{\Omega} \cdot \mathbf{\Omega}_0)^2}{2 s_\Omega{}^2}\right\}, \qquad (16)$$

where $\mathbf{x}_b = x_b(u_b, v_b)$, s_x, s_Ω and s_E are the spreads in space, angle and energy respectively, and K_Ω, K_x are normalization constants. It should be observed that in the limit as $s_x, s_\Omega, s_E \to 0$, the boundary condition (16) reduces to the Green's function boundary condition (12).

Equation (16) may now be substituted into Eq. (15) and the resulting integrals approximated by the mean value theorem and saddle point techniques discussed in [10]. The primary flux, which in this case is called the broad zero order Green's function $G^b_{jm}(\mathbf{x}, \mathbf{x}_0, \mathbf{\Omega}, \mathbf{\Omega}_0, E, E_0)$, is then given by the expression

$$G^b_{jm}(\mathbf{x}, \mathbf{x}_0, \mathbf{\Omega}, \mathbf{\Omega}_0, E, E_0) \approx \frac{P_m(\bar{E}_m)}{P_m(E)}$$

$$\cdot \frac{\delta_{jm} H[-\mathbf{\Omega} \cdot \mathbf{n}(\mathbf{x}')]}{4\pi K_x K_\Omega |\partial_u \mathbf{x}' \times \partial_v \mathbf{x}'| s_x{}^2 s_\Omega s^b_m(\rho - \rho', E_0)}$$

$$\cdot \exp\left\{-\frac{(u' - u_0)^2 + (v' - v_0)^2}{2 s_x{}^2}\right\}$$

$$\cdot \exp\left\{-\frac{(1 - \mathbf{\Omega} \cdot \mathbf{\Omega}_0)^2}{2 s_\Omega{}^2}\right\}$$

$$\cdot \exp\left\{-\frac{[E - \hat{E}_m(\rho - \rho', E_0)]^2}{2 s^b_m(\rho - \rho', E_0)^2}\right\} \qquad (17)$$

where

$$s^b_m(\rho - \rho', E_0)^2 = s_m(\rho - \rho', E_0)^2$$

$$+ \left(\frac{\tilde{S}_m[\hat{E}_m(\rho - \rho', E_0)]}{\tilde{S}_m[E_0]}\right)^2 s_E{}^2. \qquad (18)$$

2.3 The First Order Green's Function

The first order Green's function is given by the second term of the Neumann series (10) with the unit boundary condition (12). Since boundary condition (12) is a special case of the boundary condition (16), only the latter will be discussed. It may be recalled that the first generation fragment flux is determined by the formula $\mathbf{G}^1 = (\mathbf{Q} \cdot \mathbf{L} \cdot \mathbf{\Xi}) \cdot \mathbf{G}^0$. Therefore, on replacing $\mathbf{G}^0$ by the broad zero order Green's function $\mathbf{G}^b$ and expanding the result, it is found that the broad first order Green's function is given by the expression

$$G^1_{jm}(\mathbf{x}, \mathbf{x}_0, \mathbf{\Omega}, \mathbf{\Omega}_0, E, E_0) =$$

$$\cdot \int_{\rho'}^{\rho} d\rho'' \frac{P_j(E'')\tilde{S}_j(E'')}{P_j(E)\tilde{S}_j(E)} \int_{E''}^{\infty} dE_1 \int_{4\pi} d\Omega_1$$

$$\cdot \frac{H[-\mathbf{\Omega}_1 \cdot \mathbf{n}(\mathbf{x}_1')] \sigma_{jm}(\mathbf{\Omega}, \mathbf{\Omega}_1, E'', E_1)}{|\partial_u \mathbf{x}_1' \times \partial_v \mathbf{x}_1'| s_x{}^2 s_\Omega s^b_m(\rho_1'' - \rho_1', E_0)}$$

$$\cdot\frac{P_m[\overline{E}_m(\rho_1''-\rho_1',E_1)]}{4\pi K_x K_\Omega P_m[E_1]}\exp\left\{-\frac{(u_1'-u_0)^2+(v_1'-v_0)^2}{2s_x{}^2}\right\}$$

$$\cdot\exp\left\{-\frac{[E_1-\hat{E}_m(\rho_1''-\rho_1',E_0)]}{2s_m^b(\rho_1''-\rho_1',E_0)^2}^2\right\}$$

$$\cdot\exp\left\{-\frac{(1-\mathbf{\Omega}_1\cdot\mathbf{\Omega}_0)^2}{2s_\Omega{}^2}\right\}, \qquad (19)$$

where $\rho_1''=\mathbf{x}''\cdot\mathbf{\Omega}_1$, $\mathbf{x}_1'=\mathbf{x}''-(\rho_1''-\rho_1')\mathbf{\Omega}_1$ is the point where the ray through $\mathbf{x}''$, in the direction $\mathbf{\Omega}_1$ enters V, $\overline{E}_m=\overline{E}_m(\rho_1''-\rho_1',E_1)$, $\hat{E}_m=\hat{E}_m(\rho_1''-\rho_1',E_0)$ is the mean energy at depth $(\rho_1''-\rho_1')$ g/cm^2 of an m-type ion that entered the transport material with energy E_0 MeV/amu, and $s_m^b(\rho_1''-\rho_1',E_0)$ is the corresponding energy width.

The expression in Eq. (19) can be evaluated by numerical quadrature, but this is computationally expensive and therefore it is desirable to construct an analytical approximation. This can be done by making use of Taylor's theorem, the mean value theorem, and saddle point techniques as described in [10], and yields the result

$$G_{jm}^1(\mathbf{x},\mathbf{x}_0,\mathbf{\Omega},\mathbf{\Omega}_0,E,E_0)=\frac{H[-\mathbf{\Omega}\cdot\mathbf{n}(\mathbf{x}')]}{(2\pi)^{\frac{3}{2}}K_x K_\Omega s_x{}^2 s_\Omega}$$

$$\cdot\frac{1}{|\partial_u\mathbf{x}'\times\partial_v\mathbf{x}'|}\exp\left\{-\frac{(u'-u_0)^2+(v'-v_0)^2}{2s_x{}^2}\right\}$$

$$\cdot\exp\left\{-\frac{(1-\mathbf{\Omega}\cdot\mathbf{\Omega}_0)^2}{2s_\Omega{}^2}\right\}G_{jm}^{1_b}(\rho,\rho',E,E_0), \qquad (20)$$

where

$$G_{jm}^{1_b}(\rho,\rho',E,E_0)=\left[\frac{C_{jm}(\rho^*)}{2g'_{jm}(\rho^*)}\right.$$

$$\cdot\left\{\operatorname{erf}\left(\frac{g_{jm}(\rho)}{\sqrt{2}s_m^1(\rho^*)}\right)-\operatorname{erf}\left(\frac{g_{jm}(\rho')}{\sqrt{2}s_m^1(\rho^*)}\right)\right\}$$

$$-\frac{C'_{jm}(\rho^*)s_m^1(\rho^*)}{\sqrt{2\pi}g'_{jm}(\rho^*)^2}$$

$$\left.\cdot\left\{\exp\left(\frac{-g_{jm}(\rho)^2}{2s_m^1(\rho^*)^2}\right)-\exp\left(\frac{-g_{jm}(\rho')^2}{2s_m^1(\rho^*)^2}\right)\right\}\right], \qquad (21)$$

ρ^* is the root of the equation $g_{jm}(\rho'')=0$,

$$g_{jm}(\rho'')=\hat{E}_m(\rho''-\rho,E_0)-E_s[\hat{E}_m(\rho''-\rho,E_0)]$$
$$-\overline{E}_j(\rho-\rho'',E), \qquad (22)$$

$$s_m^1(\rho^*)^2=s_m^b(\rho^*-\rho',E_0)^2+\sigma^2\gamma_L$$
$$\cdot[\hat{E}_m(\rho-\rho^*,E_0)]^2\beta_L[\hat{E}_m(\rho-\rho^*,E_0)]^2, \qquad (23)$$

and

$$C_{jm}(\rho'')=\frac{\tilde{S}_j[\overline{E}_j(\rho-\rho'',E)]P_m(E_0)}{\tilde{S}_j(E)P_j(E)}$$

$$\cdot\frac{P_j[\overline{E}_j(\rho-\rho'',E)]}{P_m[\hat{E}_m(\rho-\rho'',E_0)]}\sigma_{jm}[\hat{E}_m(\rho-\rho'',E_0)]. \qquad (24)$$

3.0 DISCUSSION

To illustrate some of the theory presented above, the case in which the boundary flux consists of the *m*th component of the Galactic Cosmic Ray (GCR) spectrum is discussed. Since the GCR is isotropic and spatially uniform, boundary condition (1) takes the special form

$$\phi_j(\mathbf{x}_b,\mathbf{\Omega},E)=\delta_{jm}F_m(E), \qquad (25)$$

where the spectra $F_m(E)$ are broad functions of energy. These have been modeled by Badhwar and O'Neil (1995) and made available in tabular form at a number of solar maxima and minima between the years1958 and 1989.

In these circumstances, Eq. (15) for the primary flux takes the form

$$\phi_j^0(\mathbf{x},\mathbf{\Omega},E) = [\mathbf{G}^0 \cdot \mathbf{F}]_j(\mathbf{x},\mathbf{\Omega},E)$$
$$\approx H[-\mathbf{\Omega}\cdot\mathbf{n}(\mathbf{x}')]\frac{P_m(\overline{E}_m)}{P_m(E)}\int_{\mathbb{R}}\frac{\delta_{jm}F_m(E_1)}{\sqrt{2\pi}s_m(\rho-\rho',E_1)}$$
$$\cdot\exp\left\{-\frac{[E-\hat{E}_m(\rho-\rho',E_1)]^2}{2s_m(\rho-\rho',E_1)^2}\right\}dE_1, \qquad (26)$$

and, with the help of Taylor's theorem and the mean value theorem, may be further approximated as

$$\phi_j^0(\mathbf{x},\mathbf{\Omega},E) \approx \delta_{jm}H[-\mathbf{\Omega}\cdot\mathbf{n}(\mathbf{x}')]$$
$$\cdot\frac{P_m(\overline{E}_m)\tilde{S}_m(\overline{E}_m)}{P_m(E)\tilde{S}_m(E)}F_m(\overline{E}_m). \quad (27)$$

The first generation fragment flux also simplifies and is approximated by the expression

$$\phi_j^1(\mathbf{x},\mathbf{\Omega},E) = [\mathbf{G}^1 \cdot \mathbf{F}]_j(\mathbf{x},\mathbf{\Omega},E)$$
$$\approx H[-\mathbf{\Omega}\cdot\mathbf{n}(\mathbf{x}')]$$
$$\cdot\int_{-\infty}^{\infty}G_{jm}^{1_0}(\rho,\rho',E,E_0)F_m(E_0)dE_0 \qquad (28)$$

where $G_{jm}^{1_0}(\rho,\rho',E,E_0)$ is the special case of $G_{jm}^{1_b}(\rho,\rho',E,E_0)$ for which $s_E = 0$.

3.1 GCR on a Half-Space

In the first example to be considered, the target is an aluminum solid that occupies the half-space $V = \{(x,y,z) : z \geq 0\}$ whose boundary ∂V is the $xy-plane$. The Cartesian components of the vector $\mathbf{\Omega}$ are given by $\mathbf{\Omega} = (\sin\gamma\cos\alpha, \sin\gamma\sin\alpha, \cos\gamma)$ and are therefore completely determined by the polar angles γ and α. The measured GCR ^{56}Fe flux associated with the 1977 solar minimum [18] provides the boundary condition (25). The primary flux (27) then takes the form

$$\phi^0{}_j(\mathbf{x},\mathbf{\Omega},E) \approx \delta_{jm}\,H[\cos\gamma]$$
$$\cdot\frac{P_m(\overline{E}_m)}{P_m(E)}\frac{\tilde{S}_m(\overline{E}_m)}{\tilde{S}_m(E)}F_m(\overline{E}_m), \qquad (29)$$

where $\overline{E}_m = \overline{E}_m(z/\cos\gamma, E)$ and $0 \leq \gamma < \pi/2$.

Since the field is axisymmetric about any line parallel to the $z-axis$, the fluxes of interest are functions of z,γ, and E only. Figures 1 and 2 show the variation of the ^{56}Fe primary flux with γ and E at the points (0,0,0) and (0,0,5) respectively. Figure 3 provides a similar illustration for the first generation ^{16}O fragments at the point (0,0,5).

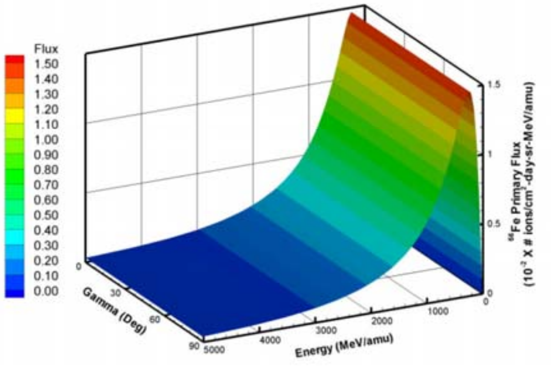


Fig 1.

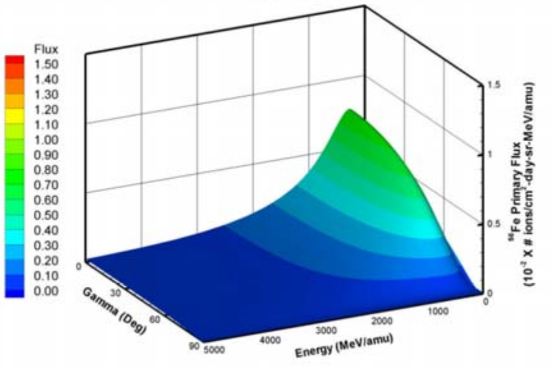


Fig 2.

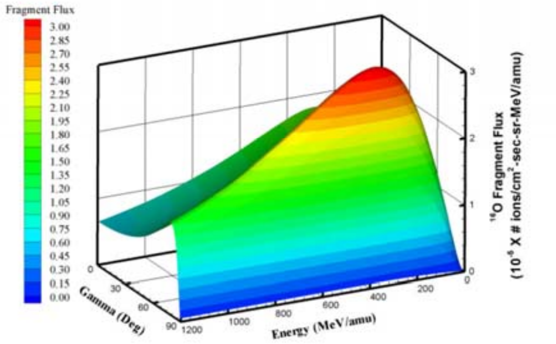


Fig 3.

3.2 GCR on a Circular Cylinder

In the second example to be considered, the target is the solid aluminum cylinder $x^2 + y^2 \leq 16^2$, $0 \leq z \leq 36$ and again the boundary condition is provided by the measured GCR ^{56}Fe flux associated with the 1977 solar minimum.

Figures 4 and 5 show the variation of the ^{56}Fe primary flux with γ and E at the points (0,0,0) and (0,0,18) respectively, where the field is axisymmetric. Similar results for the first generation ^{16}O fragment flux are shown in Figs. 6 and 7.

The remaining figures deal with the flux at the point (14,0,9) where the field is no longer axisymmetric. Figures 8 and 9 show how the ^{56}Fe primary flux varies with γ and E when $\alpha = 0°$ and $\alpha = 90°$ respectively. Figures 10 and 11 provide a similar illustration of the ^{16}O fragment flux.

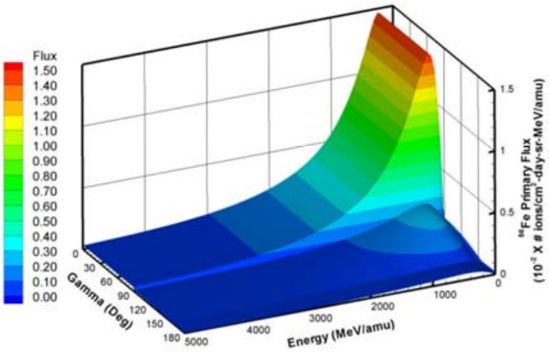


FIG 4.

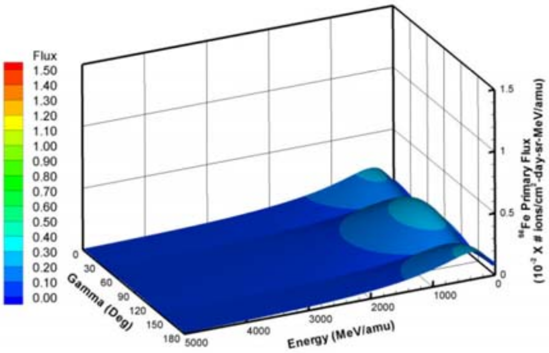


FIG 5.

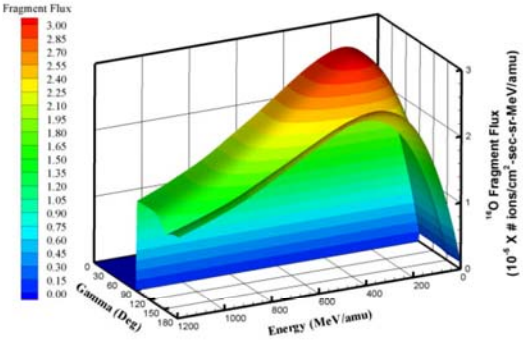


FIG 6.

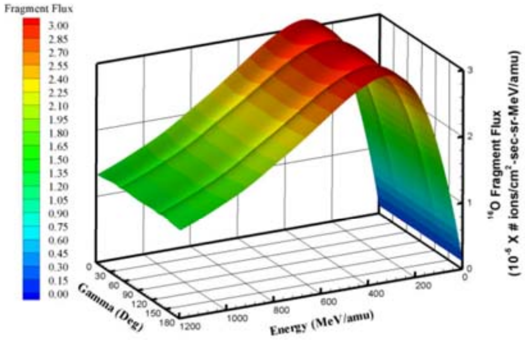


FIG 7.

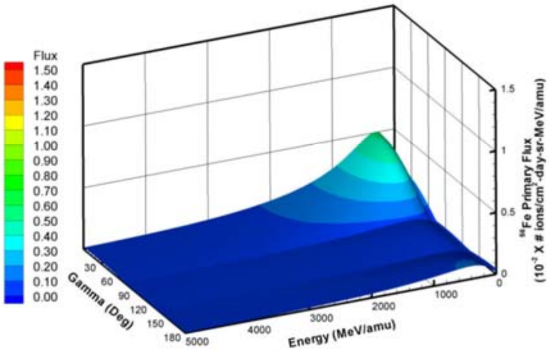


FIG 8.

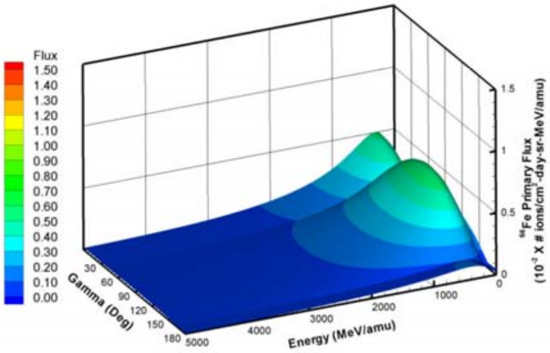


FIG 9.

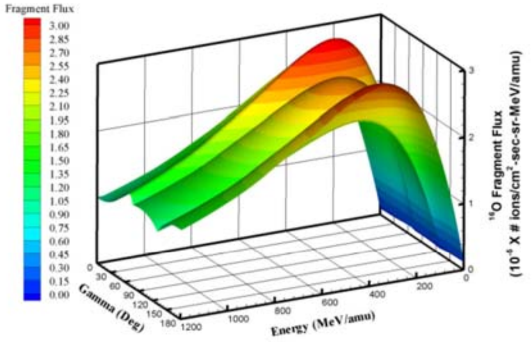


FIG 10.

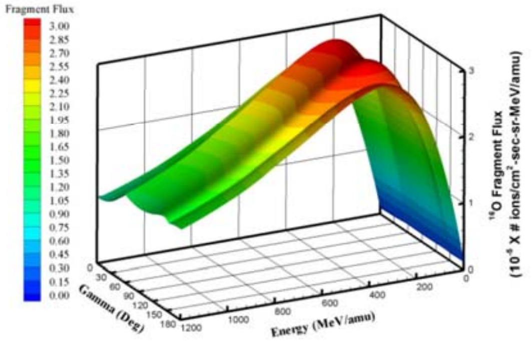


FIG 11.

4.0 CONCLUSION(S)

In this work, some progress toward the development of a fully three dimensional deterministic code for space radiation transport was discussed. Approximations were obtained for the first two terms of the Neumann series solution of the transport integral equation. The results were then illustrated by exhibiting the primary flux and the first generation ^{16}O fragment flux in a half-space, and in a circular cylinder, due to the ^{56}Fe component of the GCR.

In future work an approximation for the third Neumann series term will be obtained and the Neumann series remainder estimated by a non-perturbative technique. Predictions made by the code will then be compared with the results of laboratory beam experiments and measured space data.

5.0 REFERENCES

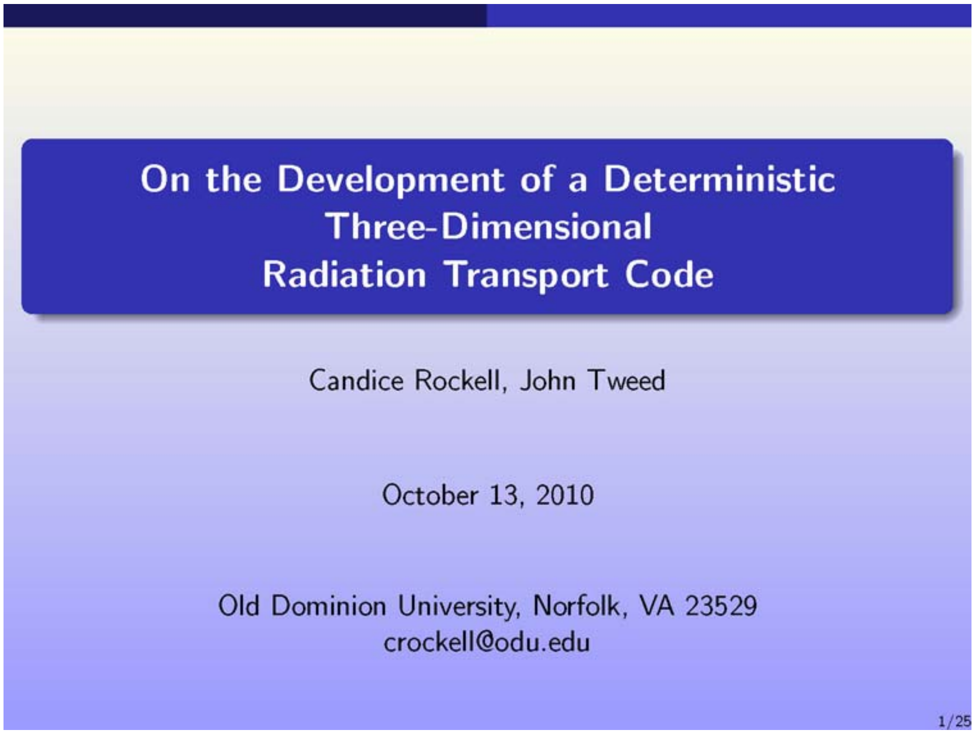
1. Committee on the Evaluation of Radiation Shielding for Space Exploration, *Managing Space Radiation Risk in the New Era of Space Exploration*, The National Academies Press, Washington, D.C. (2008).
2. J. W. Wilson, L. W. Townsend , S. L. Lamkin, and B. D. Ganapol, "A closed form solution to HZE propagation," *Radiat. Res.,* **122**, 223-228 (1990).
3. J. W. Wilson, S. L. Lamkin, H. Farhat, B. D. Ganapol, and L. W. Townsend, "A hierarchy of transport approximations for high energy heavy (HZE) ions," NASA TM-4118, National Aeronautics and Space Administration (1989).
4. J. W. Wilson, F.F. Badavi, R. C. Costen and J. L. Shinn, "Non-perturbative methods in HZE transport," NASA TP-3363, National Aeronautics and Space Administration (1993).
5. J. W. Wilson, L. W. Townsend, W. Schimmerling, G. S. Khandelwal, F. Khan, J. E. Nealy, F. A. Cucinotta, L. C. Simonsen, J. L. Shinn and J. W. Norbury, "Transport Methods and Interactions for Space Radiations," *NASA RP-1257,* National Aeronautics and Space Administration (1991).
6. J. L. Shinn, J. W. Wilson, F.F. Badavi, E. V. Benton, I. Csige, A. L. Frank and E. R. Benton, "HZE Beam Transport in Multilayered Materials," *Radiat. Meas.*, 23, 57-64 (1994).
7. J. L. Shinn, J. W. Wilson, W. Schimmerling, M. R. Shavers, J. Miller, E. V. Benton, A. L. Frank and F. F. Badavi, "A Green's function method for heavy ion beam transport," *Radiat. Environ. Biophys.*, 34, 155-159 (1995).
8. J. L. Shinn, J. W. Wilson, E. V. Benton and F.F. Badavi, "Multilayer analysis of an Iron radiation beam experiment," NASA TM-4753, National Aeronautics and Space Administration (1997).
9. S. A. Walker, J. Tweed, J. W. Wilson, F. A. Cucinotta, R. K. Tripathi, S. Blattnig, C. Zeitlin, L. Heilbronn and J. Miller,, "Validation of the HZETRN code for laboratory exposures with 1 A GeV iron ions in several targets," *Adv. Space Res.,* 35, 202-207 (2005).
10. J. Tweed, J. W. Wilson, and R. K. Tripathi, "An improved Green's function for ion beam transport," *Adv. Space Res.,* 34, 1311-1318 (2004).
11. C. J. Mertens, S. A. Walker, J. W. Wilson, R. C. Singleterry, and J. Tweed, "Coupling of Multiple Coulomb Scattering with Energy

Loss and Straggling in HZETRN," *Adv. Space Res.*, 40, 1357-1367 (2007).

12. J. Tweed, S. A. Walker, J. W. Wilson, R. K. Tripathi, F. F. Badavi, J. Miller, C. Zeitlin, and L. H. Heilbronn,, "Validation Studies of the GRNTRN Code for Radiation Transport," ICES 2007-01-3118, SAE 37th International Conference on Environmental Systems, Chicago, 2007.
13. J. Tweed, S. A. Walker, J. W. Wilson, and R. K. Tripathi, "Recent Progress in the Development of a Multi-Layer Green's Function Code for Ion Beam Transport," STAIF-2008 (Space Technology & Applications Forum), Albuquerque, NM, February 10-14, 2008. *AIP Conf. Proc.*, 969, 993-100 (2008).
14. J. W. WILSON, "Analysis of the Theory of High-Energy Ion Transport," *NASA TN D-8381,* National Aeronautics and Space Administration (1977).
15. L.W. Townsend, F. Khan and R. K. Tripathi, "Optical model analyses of 1.65A GeV argon fragmentation: Cross sections and momentum distributions," Phys. Rev. C 48, 2912-2919 (1993).
16. R. K. Tripathi , W. Townsend and F. Khan, "Role of intrinsic width in fragment momentum distributions in heavy ion collisions," *Physical Review C*, 49, 1775-1777 (1994).
17. C. Rockell, J. Tweed, S. R. Blattnig and C. J. Mertens, "Recent Developments in Three Dimensional Radiation transport using the Green's Function Technique," *Nuclear Science and Engineering,*(unpublished).
18. J. W. Wilson, F.F. Badavi, F. A. Cucinotta, J. L. Shinn, G. D. Badhwar, R. Silberberg, C. H. Tsao, L. W. Townsend and R. K. Tripathi, HZETRN: Description of a Free-Space Ion and Nucleon Transport and Shielding Computer Program, NASA TP-3495, National Aeronautics and Space Administration, (1995).

6.0 ACKNOWLEDGMENTS

The author (C.R.) wishes to acknowledge the NASA support she received for this work under the Graduate Student Researchers Program (NASA NNX09AJ06H) and the Virginia Space Grant Consortium.



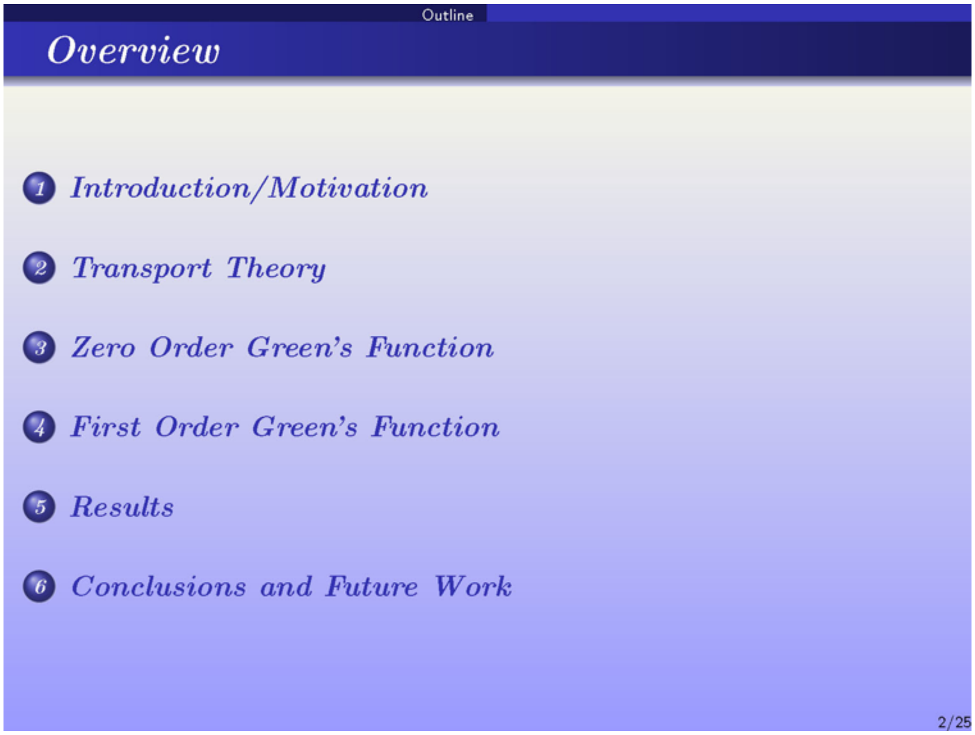
On the Development of a Deterministic Three-Dimensional Radiation Transport Code

Candice Rockell, John Tweed

October 13, 2010

Old Dominion University, Norfolk, VA 23529
crockell@odu.edu

1/25

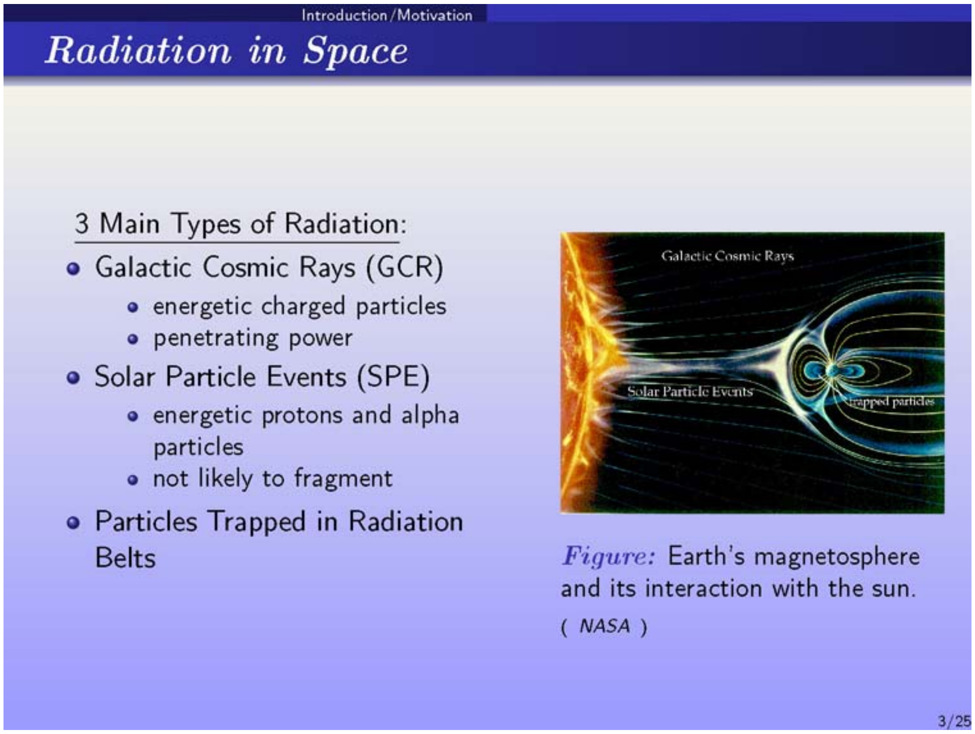


Outline

Overview

1. *Introduction/Motivation*
2. *Transport Theory*
3. *Zero Order Green's Function*
4. *First Order Green's Function*
5. *Results*
6. *Conclusions and Future Work*

2/25



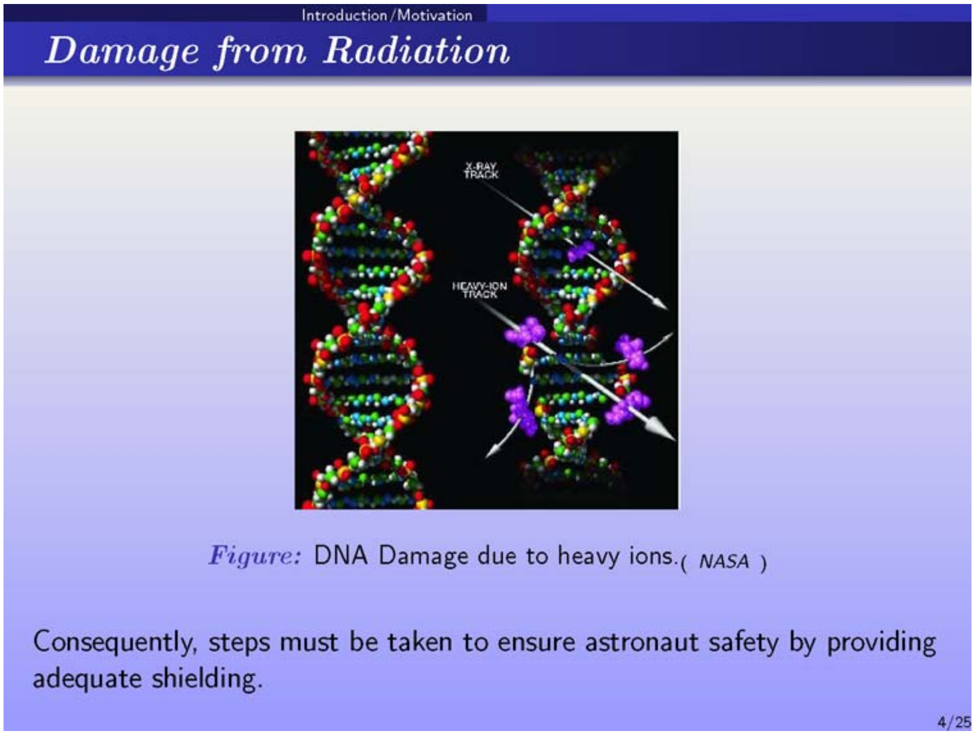
Radiation in Space

3 Main Types of Radiation:

- Galactic Cosmic Rays (GCR)
  - energetic charged particles
  - penetrating power
- Solar Particle Events (SPE)
  - energetic protons and alpha particles
  - not likely to fragment
- Particles Trapped in Radiation Belts

Galactic Cosmic Rays
Solar Particle Events
Trapped particles

Figure: Earth's magnetosphere and its interaction with the sun. (*NASA*)

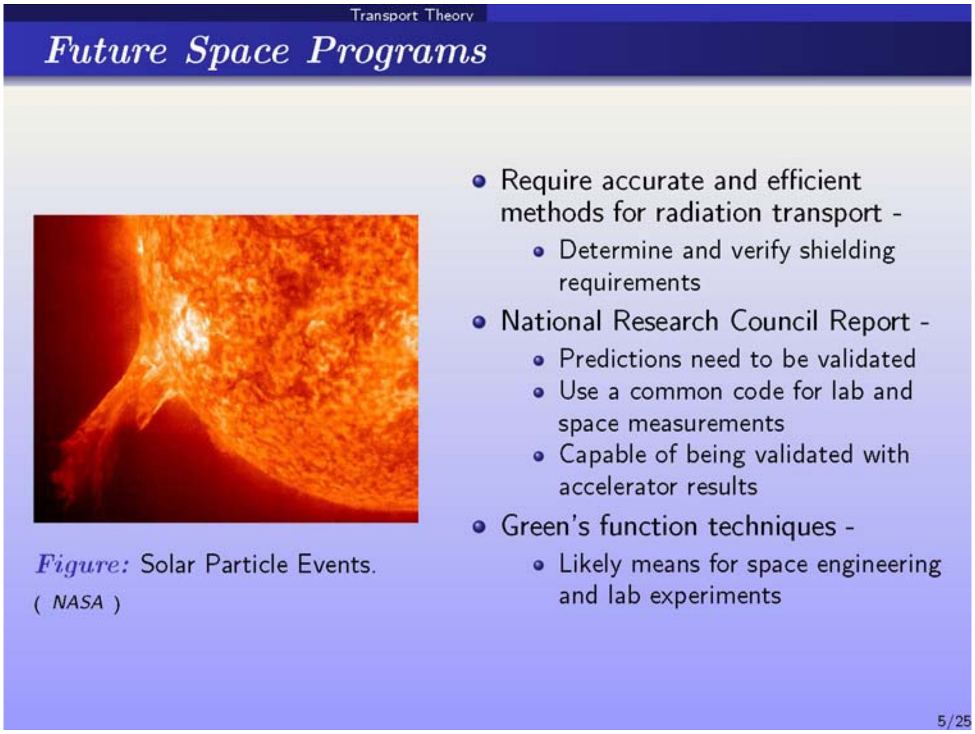


Damage from Radiation

X-RAY TRACK
HEAVY-ION TRACK

Figure: DNA Damage due to heavy ions.(*NASA*)

Consequently, steps must be taken to ensure astronaut safety by providing adequate shielding.

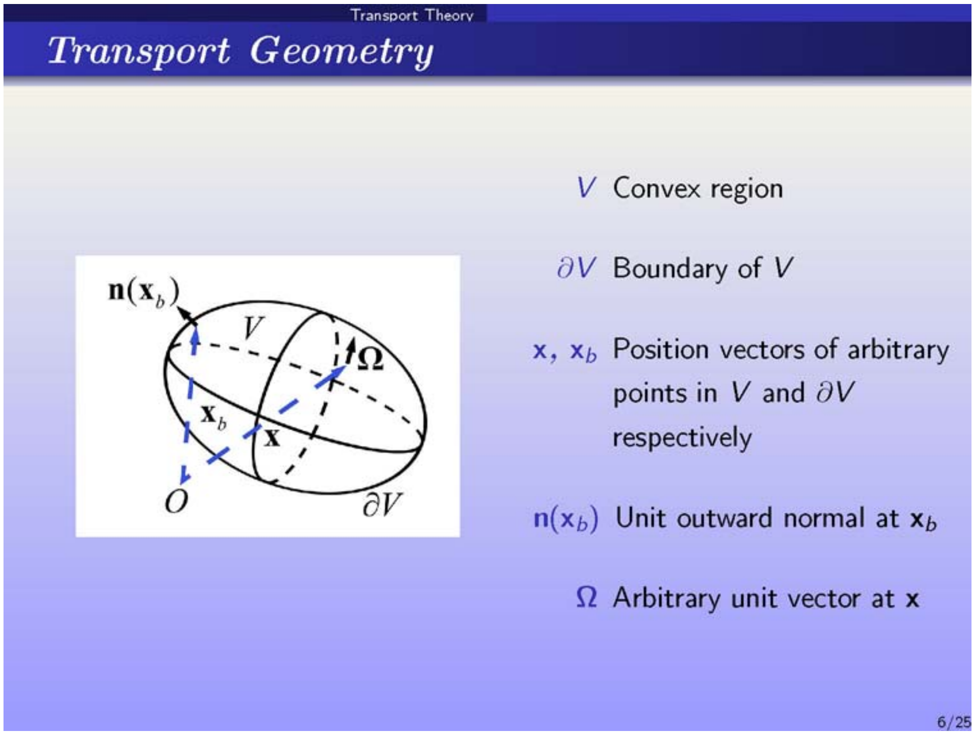


Future Space Programs

- Require accurate and efficient methods for radiation transport -
  - Determine and verify shielding requirements
- National Research Council Report -
  - Predictions need to be validated
  - Use a common code for lab and space measurements
  - Capable of being validated with accelerator results
- Green's function techniques -
  - Likely means for space engineering and lab experiments

Figure: Solar Particle Events.

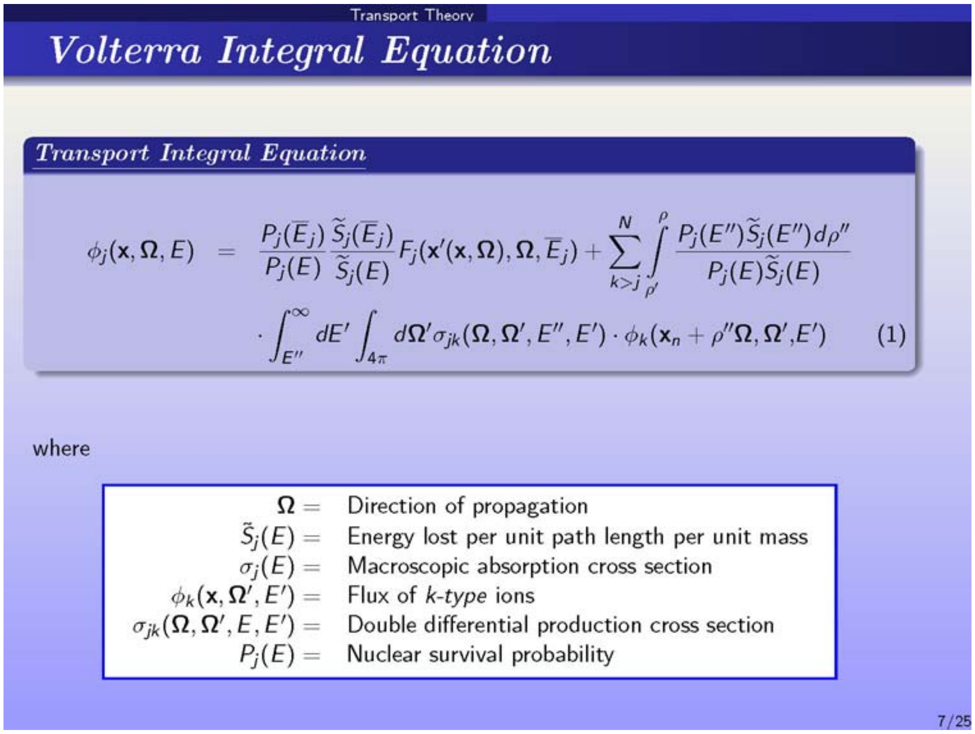
(*NASA*)



Transport Geometry

$\mathbf{n}(\mathbf{x}_b)$
V
$\mathbf{\Omega}$
$\mathbf{x}_b$
$\mathbf{x}$
O
∂V

V	Convex region
∂V	Boundary of V
$\mathbf{x}$, $\mathbf{x}_b$	Position vectors of arbitrary points in V and ∂V respectively
$\mathbf{n}(\mathbf{x}_b)$	Unit outward normal at $\mathbf{x}_b$
Ω	Arbitrary unit vector at $\mathbf{x}$



Transport Theory

Volterra Integral Equation

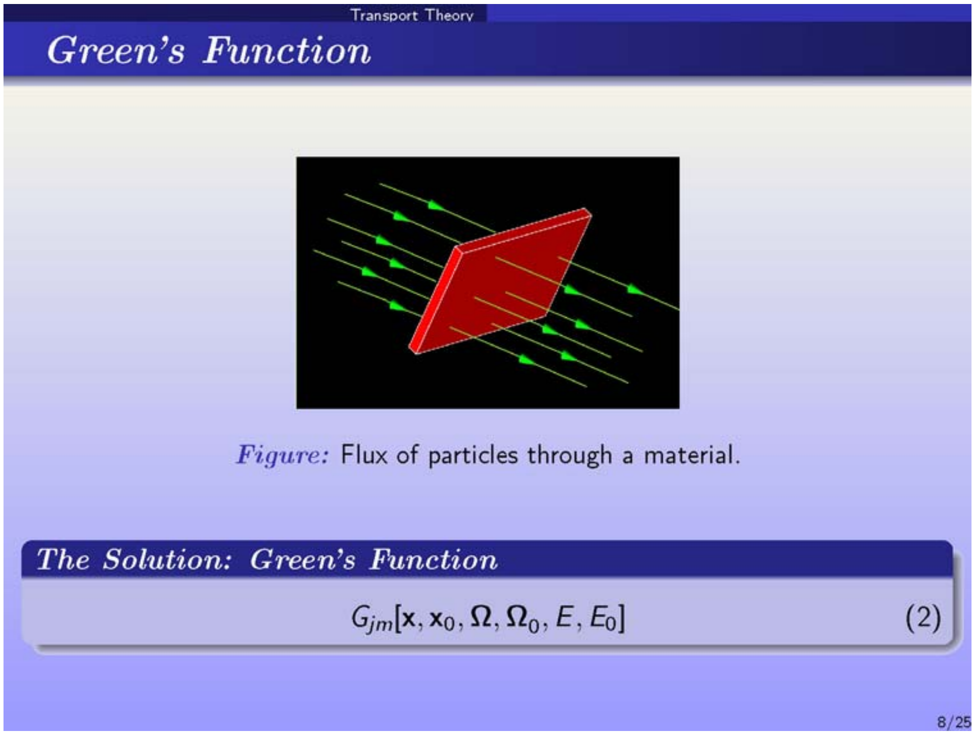
Transport Integral Equation

$$\phi_j(\mathbf{x},\mathbf{\Omega},E) = \frac{P_j(\overline{E}_j)}{P_j(E)}\frac{\widetilde{S}_j(\overline{E}_j)}{\widetilde{S}_j(E)}F_j(\mathbf{x}'(\mathbf{x},\mathbf{\Omega}),\mathbf{\Omega},\overline{E}_j) + \sum_{k>j}^{N}\int_{\rho'}^{\rho}\frac{P_j(E'')\widetilde{S}_j(E'')d\rho''}{P_j(E)\widetilde{S}_j(E)}$$
$$\cdot\int_{E''}^{\infty}dE'\int_{4\pi}d\mathbf{\Omega}'\sigma_{jk}(\mathbf{\Omega},\mathbf{\Omega}',E'',E')\cdot\phi_k(\mathbf{x}_n+\rho''\mathbf{\Omega},\mathbf{\Omega}',E') \qquad (1)$$

where

$\mathbf{\Omega} =$	Direction of propagation
$\tilde{S}_j(E) =$	Energy lost per unit path length per unit mass
$\sigma_j(E) =$	Macroscopic absorption cross section
$\phi_k(\mathbf{x},\mathbf{\Omega}',E') =$	Flux of *k-type* ions
$\sigma_{jk}(\mathbf{\Omega},\mathbf{\Omega}',E,E') =$	Double differential production cross section
$P_j(E) =$	Nuclear survival probability

7/25



Transport Theory

Green's Function

Figure: Flux of particles through a material.

The Solution: Green's Function

$$G_{jm}[\mathbf{x},\mathbf{x}_0,\mathbf{\Omega},\mathbf{\Omega}_0,E,E_0] \qquad (2)$$

8/25

Transport Theory

Neumann Series Solution

This Neumann series can be expressed as

$$\mathbf{\Phi} = [\mathbf{G}^0 + \mathbf{G}^1 + \mathbf{G}^2 + \mathbf{G}^3 + \ldots] \cdot \mathbf{F} \tag{3}$$

Figure: Atmospheric air shower.(*Pierre Auger Observatory Team*)

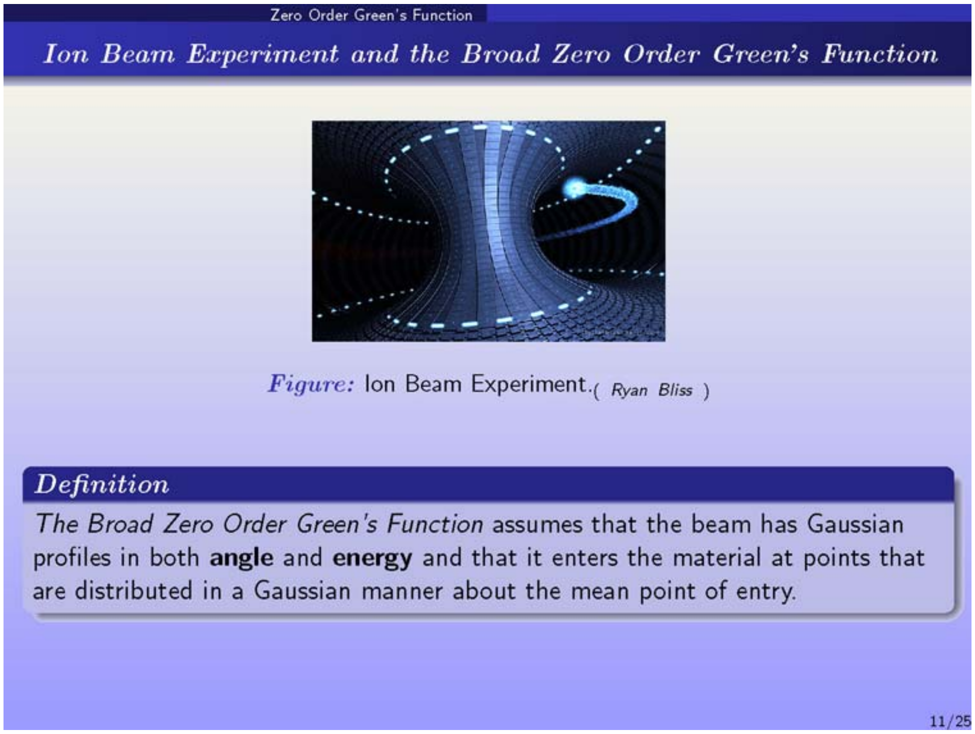
Zero Order Green's Function

Primary Flux

The primary flux can be obtained by

$$\begin{aligned} \phi_j^0(\mathbf{x}, \mathbf{\Omega}, E) &= [\mathbf{G^0} \cdot \mathbf{F}]_j(\mathbf{x}, \mathbf{\Omega}, E) \\ &= \int_{\partial V} d\mathbf{x}_0 \int_{4\pi} d\mathbf{\Omega}_0 \int_E^{\infty} dE_0 \\ &\quad \cdot G_{jj}^0(\mathbf{x}, \mathbf{x}_0, \mathbf{\Omega}, \mathbf{\Omega}_0, E, E_0) F_j(\mathbf{x}_0, \mathbf{\Omega}_0, E_0), \end{aligned} \tag{4}$$

which needs to be evaluated numerically for some situations.



Zero Order Green's Function

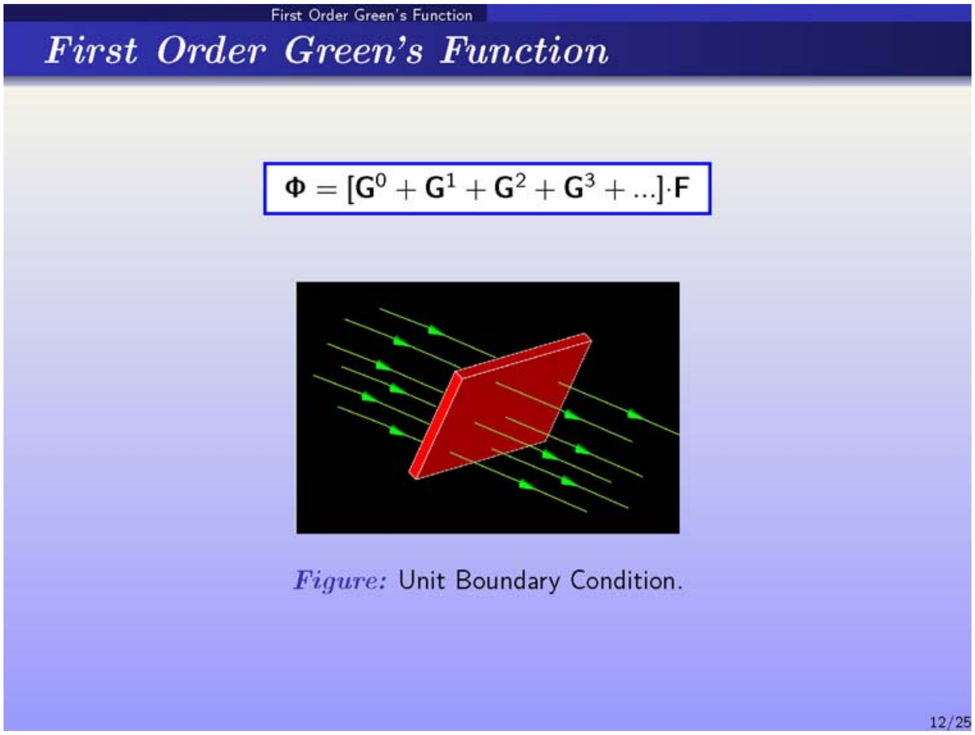
Ion Beam Experiment and the Broad Zero Order Green's Function

Figure: Ion Beam Experiment.(*Ryan Bliss*)

Definition

The Broad Zero Order Green's Function assumes that the beam has Gaussian profiles in both **angle** and **energy** and that it enters the material at points that are distributed in a Gaussian manner about the mean point of entry.

11/25



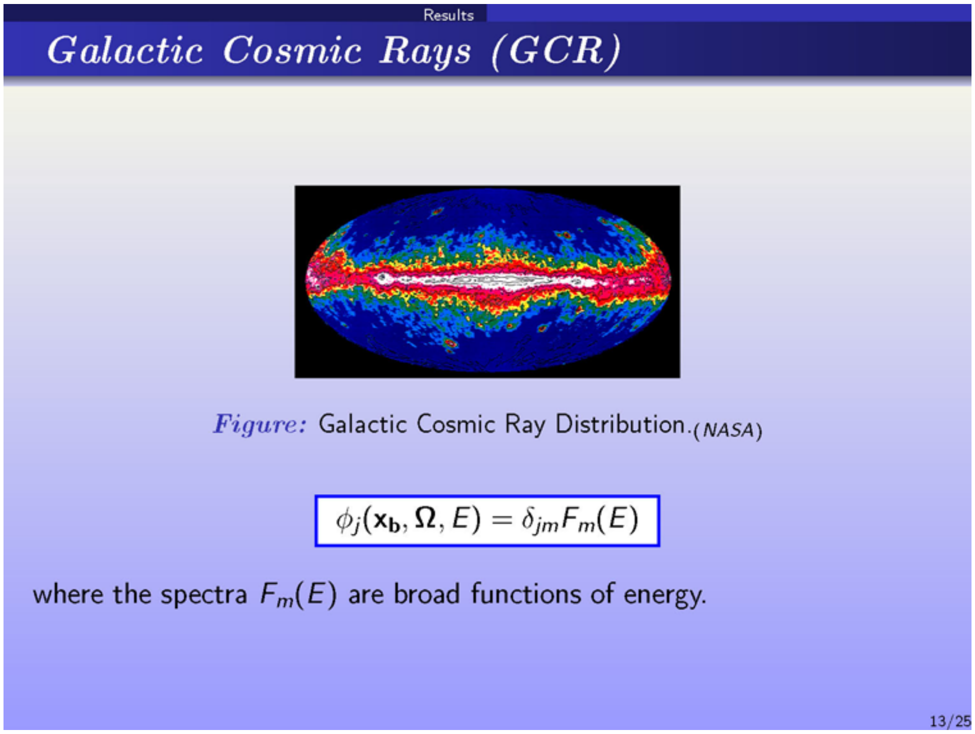
First Order Green's Function

First Order Green's Function

$$\boxed{\mathbf{\Phi} = [\mathbf{G}^0 + \mathbf{G}^1 + \mathbf{G}^2 + \mathbf{G}^3 + \ldots]\cdot\mathbf{F}}$$

Figure: Unit Boundary Condition.

12/25



Results

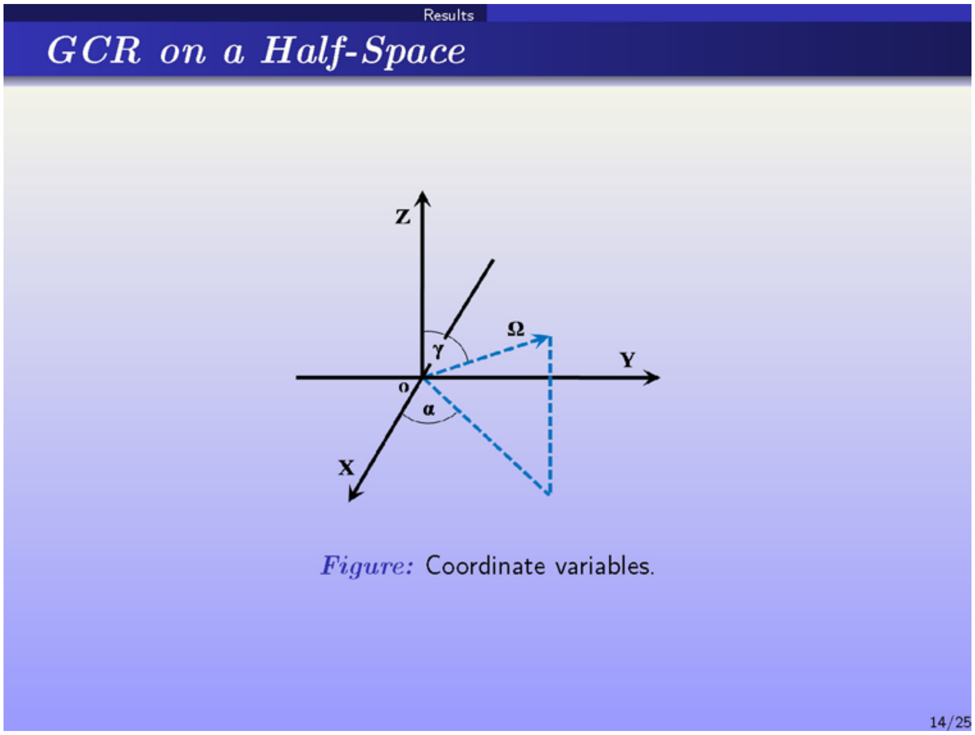
Galactic Cosmic Rays (GCR)

Figure: Galactic Cosmic Ray Distribution.$_{(NASA)}$

$$\boxed{\phi_j(\mathbf{x_b}, \mathbf{\Omega}, E) = \delta_{jm} F_m(E)}$$

where the spectra $F_m(E)$ are broad functions of energy.

13/25

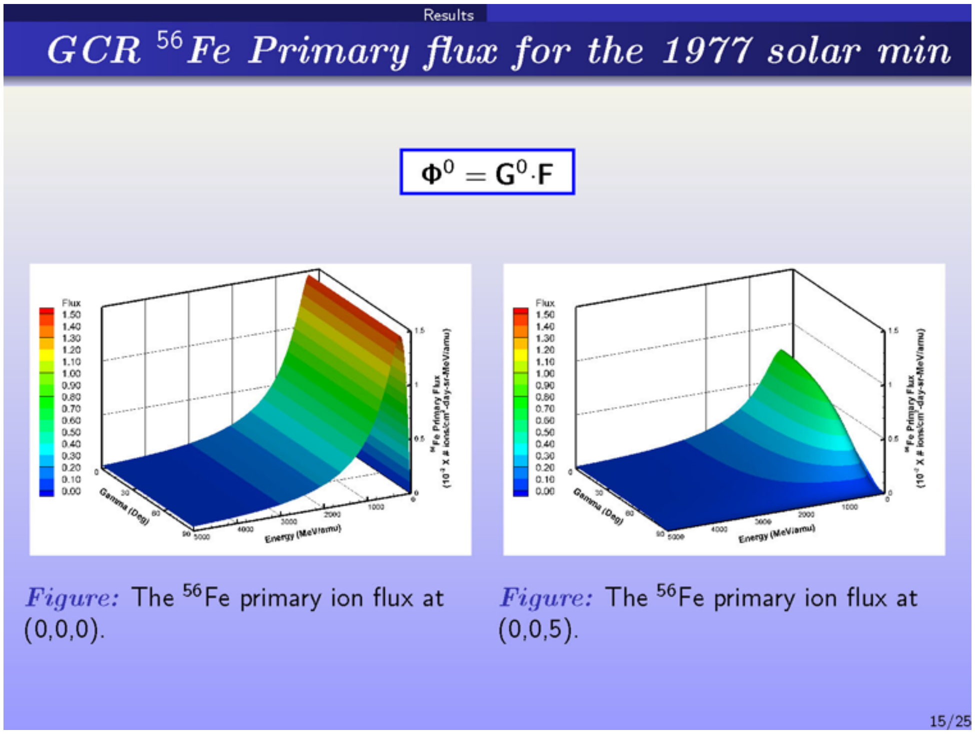


Results

GCR on a Half-Space

Figure: Coordinate variables.

14/25

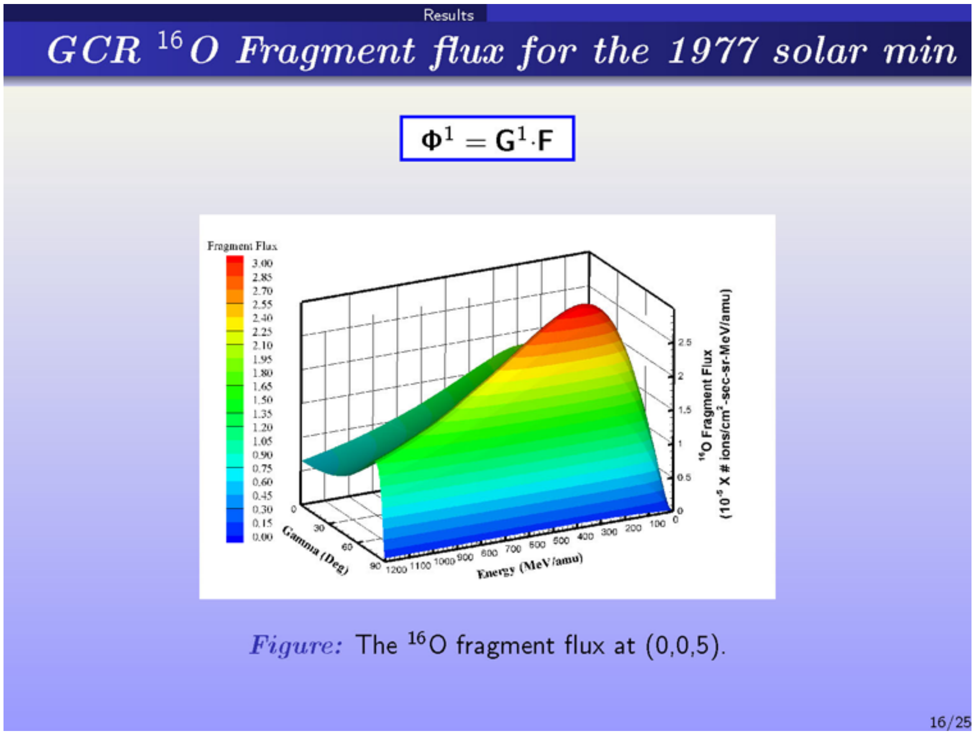


GCR ^{56}Fe Primary flux for the 1977 solar min

$$\boxed{\Phi^0 = G^0 \cdot F}$$

Figure: The ^{56}Fe primary ion flux at (0,0,0).

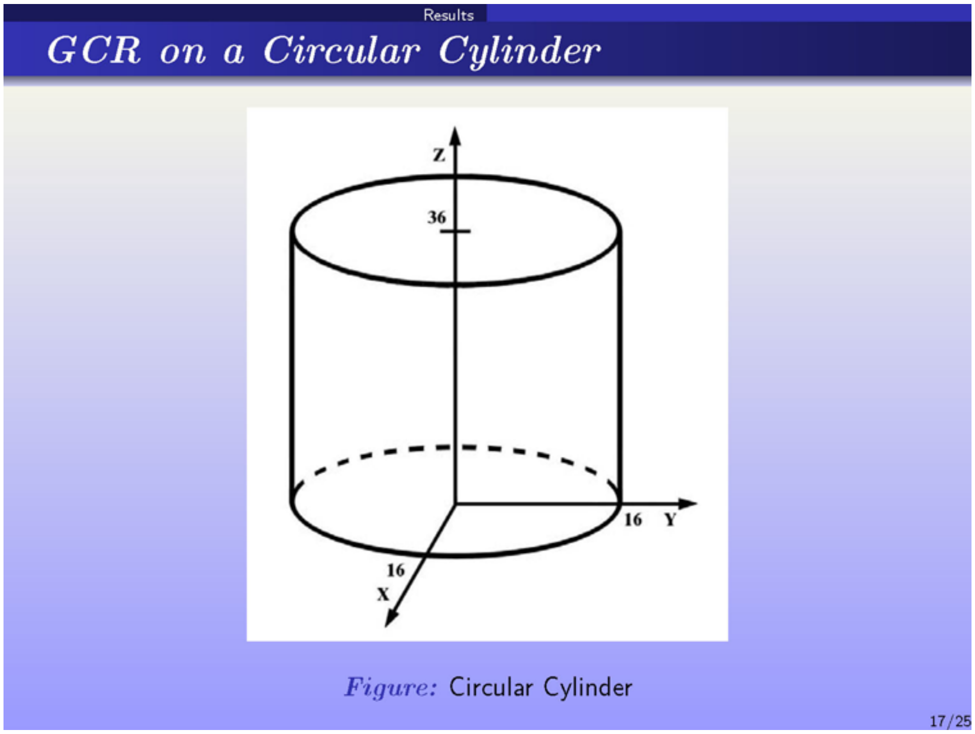
Figure: The ^{56}Fe primary ion flux at (0,0,5).



GCR ^{16}O Fragment flux for the 1977 solar min

$$\boxed{\Phi^1 = G^1 \cdot F}$$

Figure: The ^{16}O fragment flux at (0,0,5).

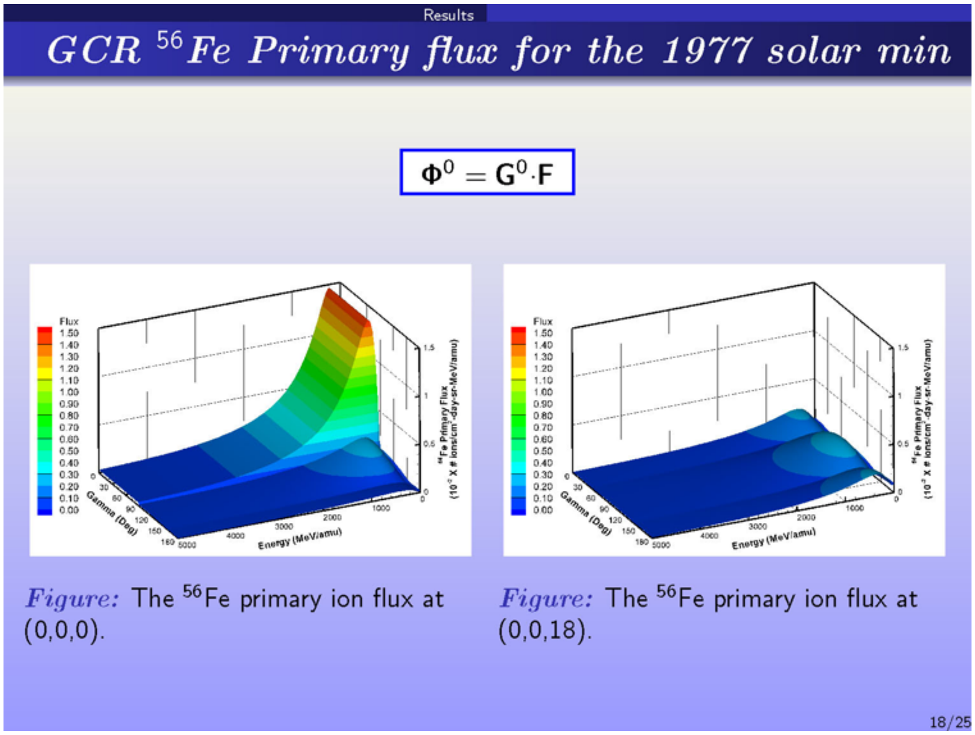


Results

GCR on a Circular Cylinder

Figure: Circular Cylinder

17/25



Results

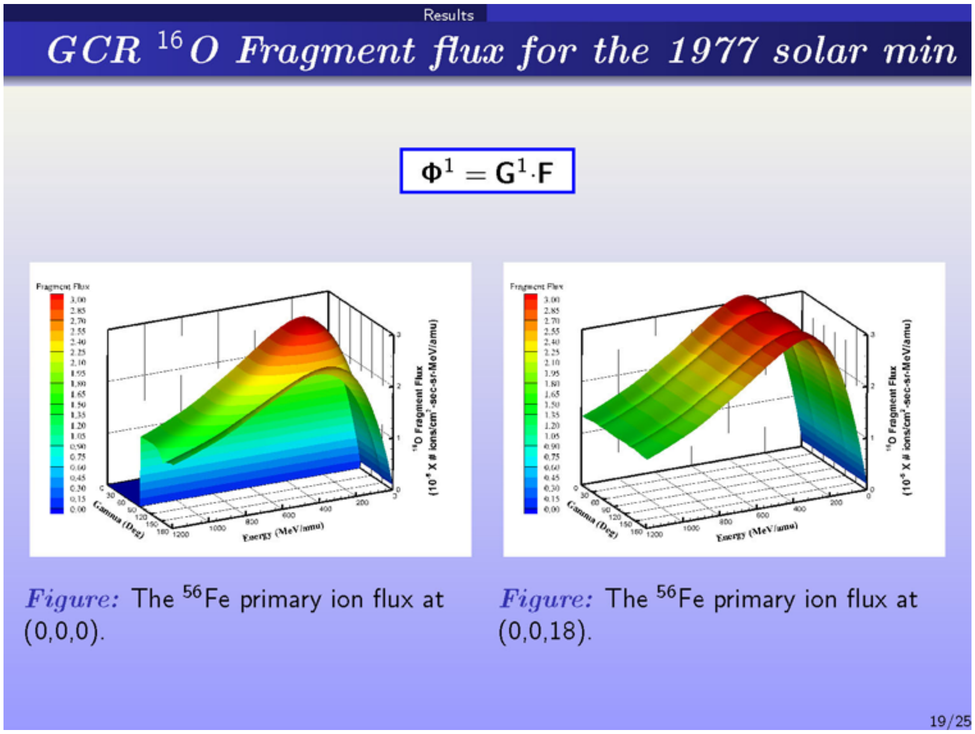
GCR ^{56}Fe Primary flux for the 1977 solar min

$$\boxed{\Phi^0 = G^0 \cdot F}$$

Figure: The ^{56}Fe primary ion flux at (0,0,0).

Figure: The ^{56}Fe primary ion flux at (0,0,18).

18/25



Results

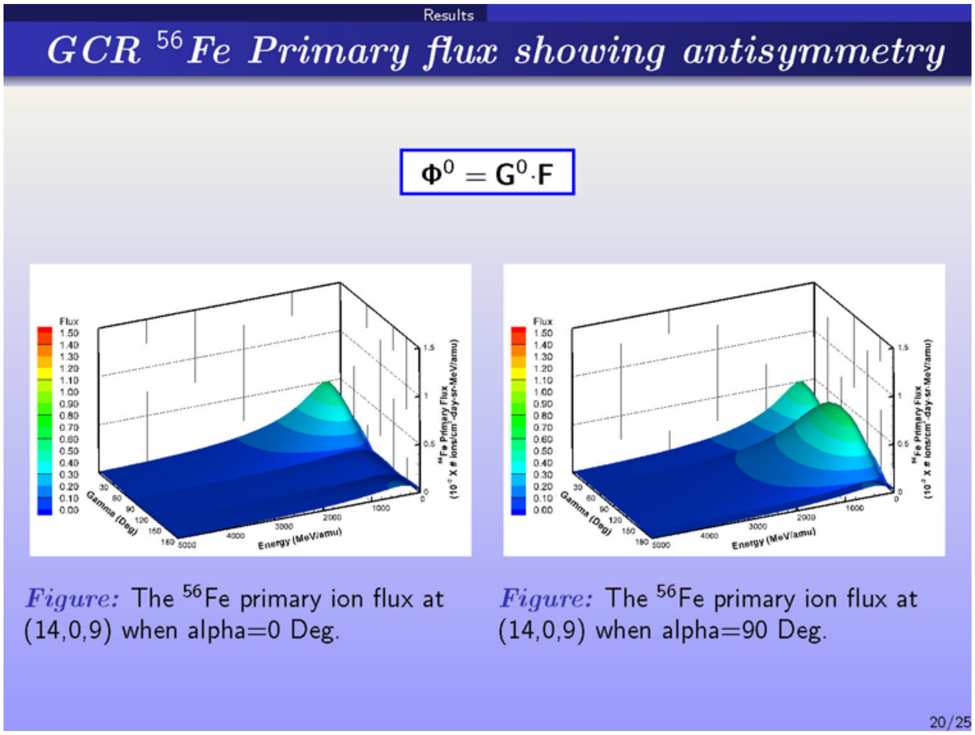
GCR ^{16}O Fragment flux for the 1977 solar min

$$\boxed{\Phi^1 = G^1 \cdot F}$$

Figure: The ^{56}Fe primary ion flux at (0,0,0).

Figure: The ^{56}Fe primary ion flux at (0,0,18).

19/25



Results

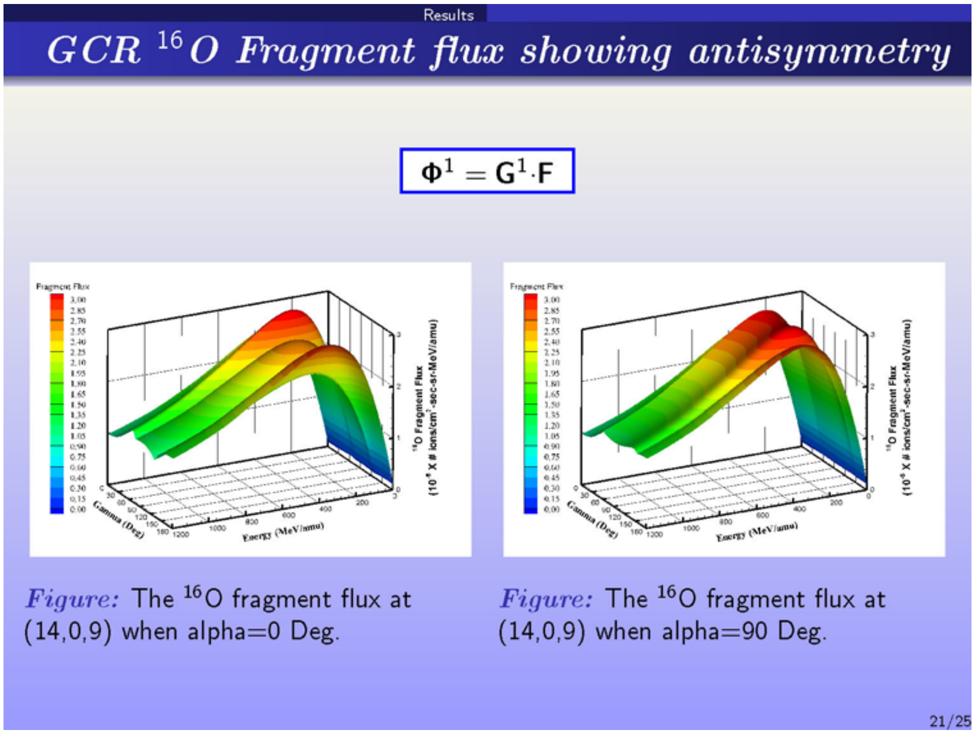
GCR ^{56}Fe Primary flux showing antisymmetry

$$\boxed{\Phi^0 = G^0 \cdot F}$$

Figure: The ^{56}Fe primary ion flux at (14,0,9) when alpha=0 Deg.

Figure: The ^{56}Fe primary ion flux at (14,0,9) when alpha=90 Deg.

20/25

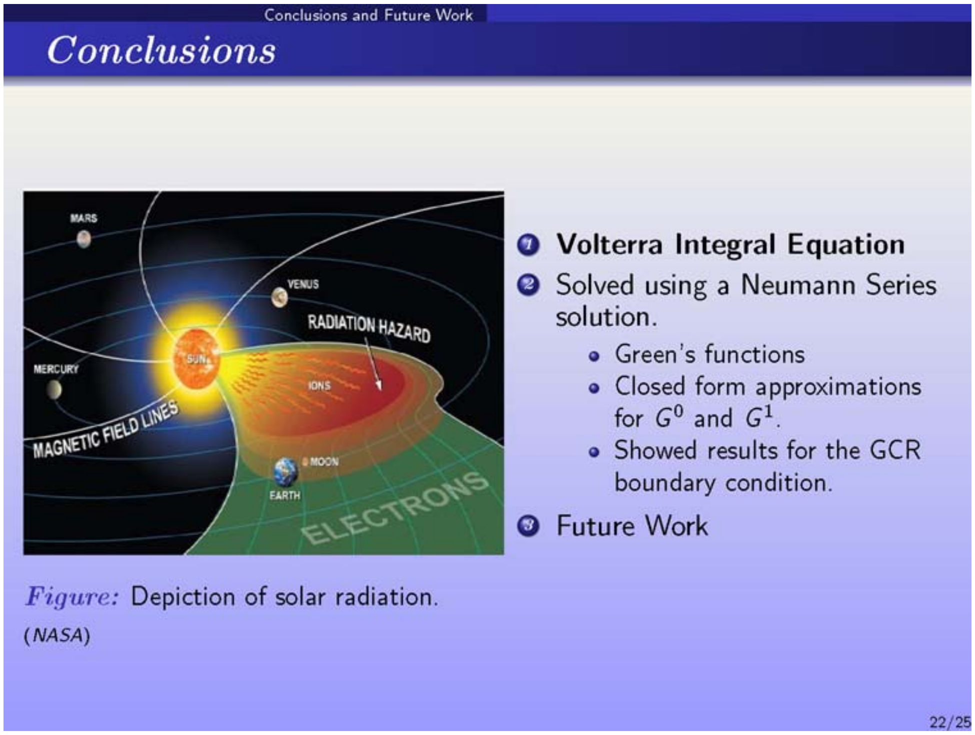


GCR ^{16}O Fragment flux showing antisymmetry

$$\boxed{\Phi^1 = G^1 \cdot F}$$

Figure: The ^{16}O fragment flux at (14,0,9) when alpha=0 Deg.

Figure: The ^{16}O fragment flux at (14,0,9) when alpha=90 Deg.



Conclusions

1. **Volterra Integral Equation**
2. Solved using a Neumann Series solution.
   - Green's functions
   - Closed form approximations for G^0 and G^1.
   - Showed results for the GCR boundary condition.
3. Future Work

Figure: Depiction of solar radiation.

(*NASA*)

Acknowledgements

I would like to acknowledge the NASA support that I received for this work under the Graduate Student Research Program (NASA NNX09AJ06H) and the Virginia Space Grant Consortium.

GSRP
Graduate Student Researchers Program
VIRGINIA SPACE GRANT CONSORTIUM

References

Committee on the Evaluation of Radiation Shielding for Space Exploration, Managing Space Radiation Risk in the New Era of Space Exploration, The National Academies Press, Washington, D.C., 2008.

Mertens, C., Wilson, J.W., et al., Coupling of Multiple Coulomb Scattering with Energy Loss and Straggling in HZETRN, Advances in Space Research 40, 1357-1367, 2007.

Miroshnichenko, Leonty I. Radiation Hazard in Space, Kluwer Academic Publishers, Boston, MA, 2003.

Shinn, J.L., et al, HZE Beam Transport in Multilayered Materials, Radiation Measurements 23, 57-64, 1994.

Shinn, J.L., et al, A Green's function method for heavy ion beam transport, Radiat. Environ. Biophys 34, 155-159, 1995.

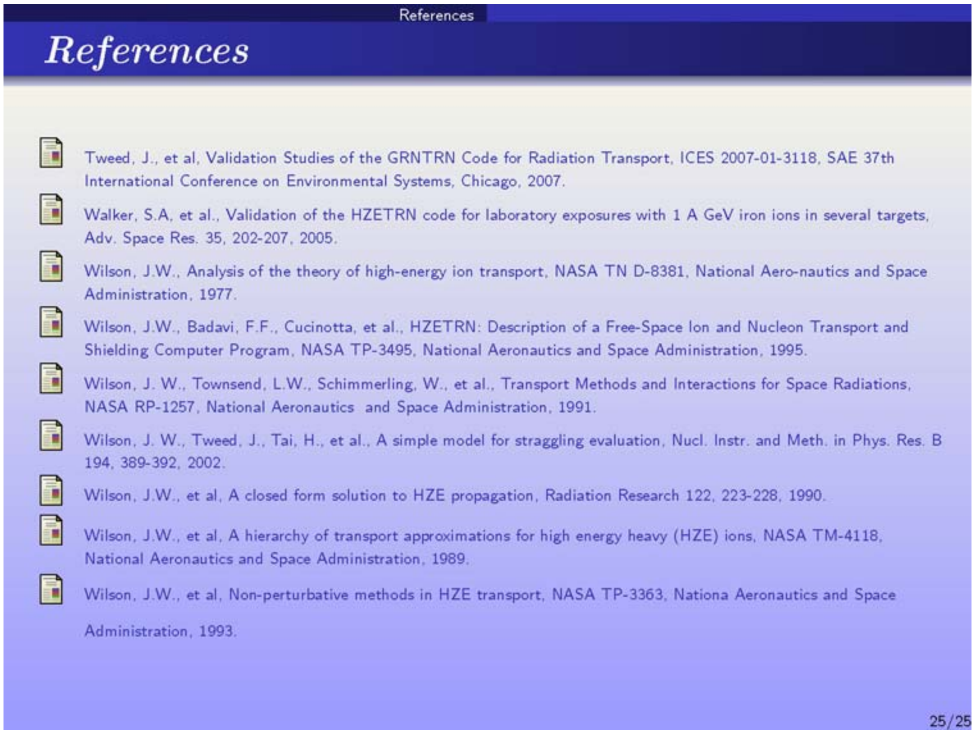
Shinn, J.L., et al, Multilayer analysis of an Iron radiation beam experiment, NASA TM-4753, National Aeronautics and Space Administration, 1997.

Townsend, L.W., Khan, F., and Tripathi, R.K., Optical model analyses of 1.65A GeV argon fragmentation: Cross sections and momentum distributions, Phys. Rev. C 48, 2912-2919, 1993.

Tripathi, R.K., Townsend, L.W., Role of intrinsic width in fragment momentum distributions in heavy ion collisions, Phys. Rev. C 48, R1775-R1777, 1994.

Tweed, J., Wilson, J. W., Tripathi, R. K., An improved Green's function for ion beam transport, Advances in Space Research 34, 1311-1318, 2004.

Tweed, J., et al., Recent Progress in the Development of a Multi-Layer Green's Fucntion Code for Ion Beam Transport, STAIF-2008 (Space Technology and Applications Forum), Albuquerque, NM, February 10-14, 2008. AIP Conf. Proc. 969, 993-1000, 2008.



References

Tweed, J., et al, Validation Studies of the GRNTRN Code for Radiation Transport, ICES 2007-01-3118, SAE 37th International Conference on Environmental Systems, Chicago, 2007.

Walker, S.A, et al., Validation of the HZETRN code for laboratory exposures with 1 A GeV iron ions in several targets, Adv. Space Res. 35, 202-207, 2005.

Wilson, J.W., Analysis of the theory of high-energy ion transport, NASA TN D-8381, National Aero-nautics and Space Administration, 1977.

Wilson, J.W., Badavi, F.F., Cucinotta, et al., HZETRN: Description of a Free-Space Ion and Nucleon Transport and Shielding Computer Program, NASA TP-3495, National Aeronautics and Space Administration, 1995.

Wilson, J. W., Townsend, L.W., Schimmerling, W., et al., Transport Methods and Interactions for Space Radiations, NASA RP-1257, National Aeronautics and Space Administration, 1991.

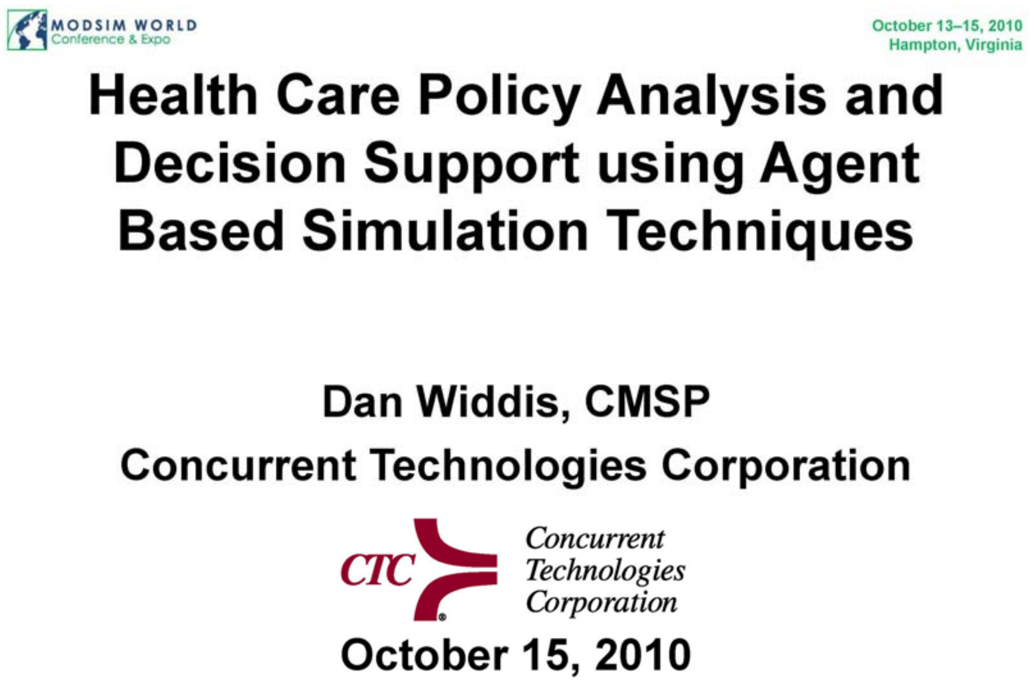
Wilson, J. W., Tweed, J., Tai, H., et al., A simple model for straggling evaluation, Nucl. Instr. and Meth. in Phys. Res. B 194, 389-392, 2002.

Wilson, J.W., et al, A closed form solution to HZE propagation, Radiation Research 122, 223-228, 1990.

Wilson, J.W., et al, A hierarchy of transport approximations for high energy heavy (HZE) ions, NASA TM-4118, National Aeronautics and Space Administration, 1989.

Wilson, J.W., et al, Non-perturbative methods in HZE transport, NASA TP-3363, Nationa Aeronautics and Space Administration, 1993.

2.21 On the Development of a Deterministic Three-Dimensional Radiation Transport Code


MODSIM WORLD Conference & Expo

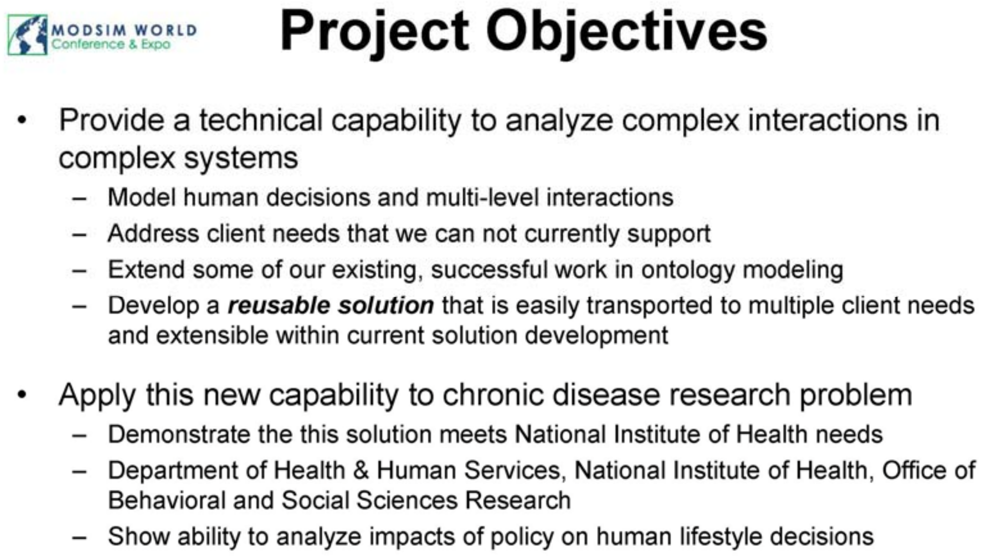
October 13–15, 2010
Hampton, Virginia

Health Care Policy Analysis and Decision Support using Agent Based Simulation Techniques

Dan Widdis, CMSP
Concurrent Technologies Corporation

CTC Concurrent Technologies Corporation

October 15, 2010


MODSIM WORLD Conference & Expo

Project Objectives

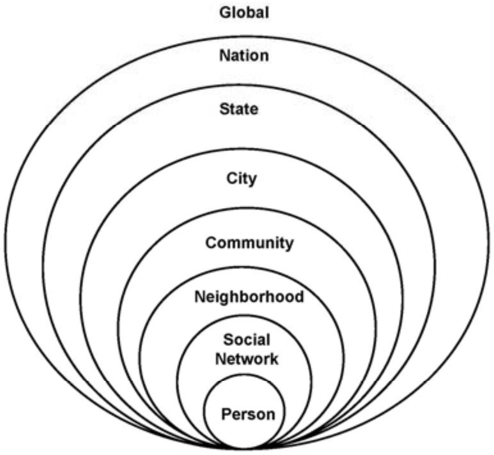
- Provide a technical capability to analyze complex interactions in complex systems
  - Model human decisions and multi-level interactions
  - Address client needs that we can not currently support
  - Extend some of our existing, successful work in ontology modeling
  - Develop a ***reusable solution*** that is easily transported to multiple client needs and extensible within current solution development
- Apply this new capability to chronic disease research problem
  - Demonstrate the this solution meets National Institute of Health needs
  - Department of Health & Human Services, National Institute of Health, Office of Behavioral and Social Sciences Research
  - Show ability to analyze impacts of policy on human lifestyle decisions

R & D Approach

- Research health care policy areas and integrate specific focus area data into usable format
  - Based on human decision model
  - Initial focus on human smoking decisions across multiple factors
- Develop ontology models
  - Human decisions
  - Human environmental entities and relationships
  - Cross-domain ontology model of interactions between humans, environment, decisions, and policy
- Develop Agent Based Model (ABM) and methods
  - Document a Design of Experiment (DOE) method
  - Document an approach to analyze ABM output

Health Care Research

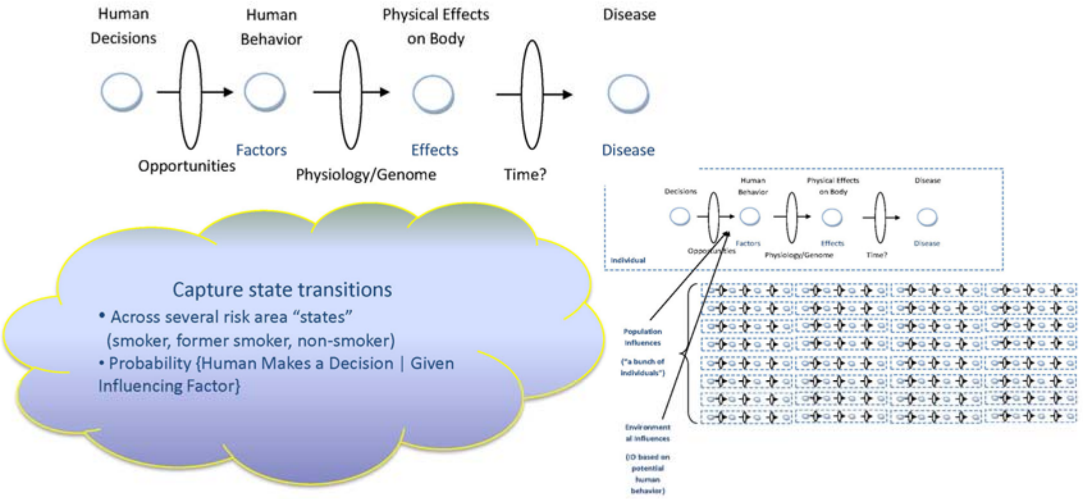
- Extensive research has been done on individual social risk factors that lead to disease
- Risk factors do not act independently
- This research allows understanding of inter-relationships between environmental influences and social influences on human decisions across many risk factors
- Enable inclusion of many risk factors across many “layers”
- Initial focus on smoking risks



Health Care Policy Focus Area

MODSIM WORLD Conference & Expo

- impact on human decisions



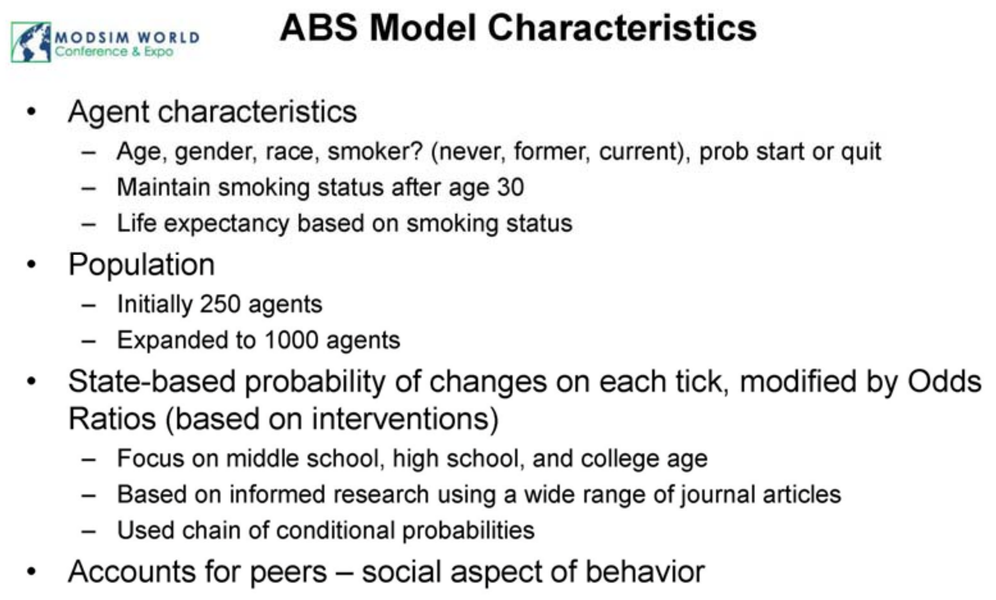
Capture state transitions

- Across several risk area "states" (smoker, former smoker, non-smoker)
- Probability {Human Makes a Decision | Given Influencing Factor}

Agent Based Simulation

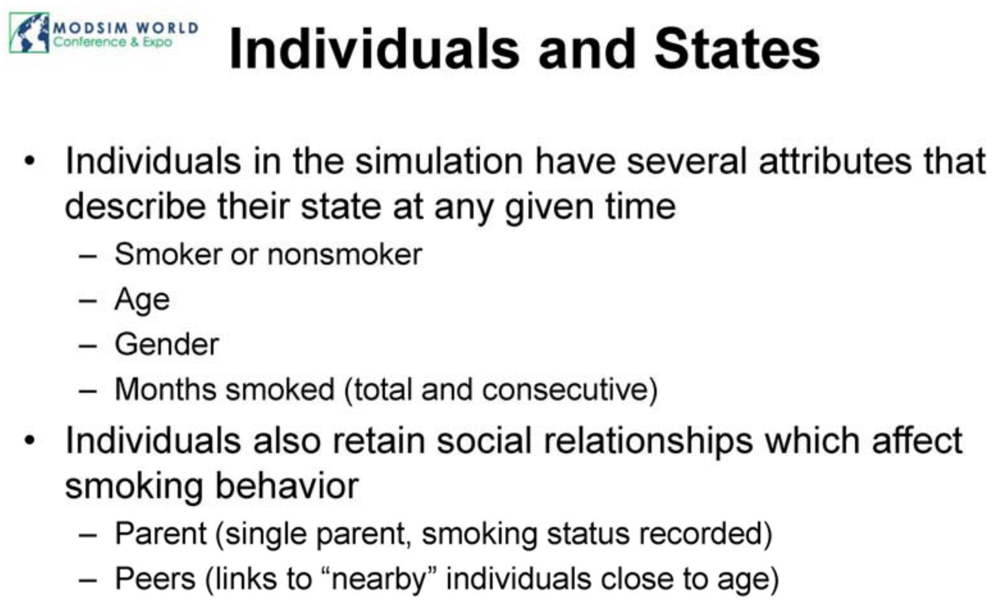
MODSIM WORLD Conference & Expo

- Collection of autonomous decision-making entities (agents)
  - NOT intelligent agents or secret agents
- Allows us to model complexity – multiple system layers and complex interactions
- Discovers "emergent phenomena"
- Becomes a data source for advanced research
- Requires sophisticated methods for:
  - Efficient experimental design
  - Data mining
- Requires computational power



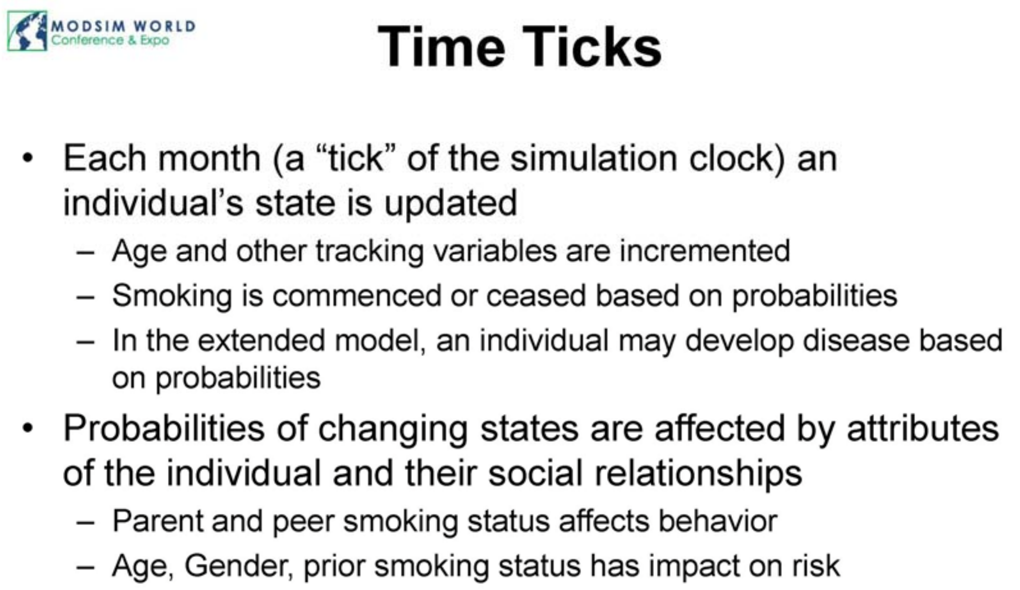
ABS Model Characteristics

- Agent characteristics
  - Age, gender, race, smoker? (never, former, current), prob start or quit
  - Maintain smoking status after age 30
  - Life expectancy based on smoking status
- Population
  - Initially 250 agents
  - Expanded to 1000 agents
- State-based probability of changes on each tick, modified by Odds Ratios (based on interventions)
  - Focus on middle school, high school, and college age
  - Based on informed research using a wide range of journal articles
  - Used chain of conditional probabilities
- Accounts for peers – social aspect of behavior



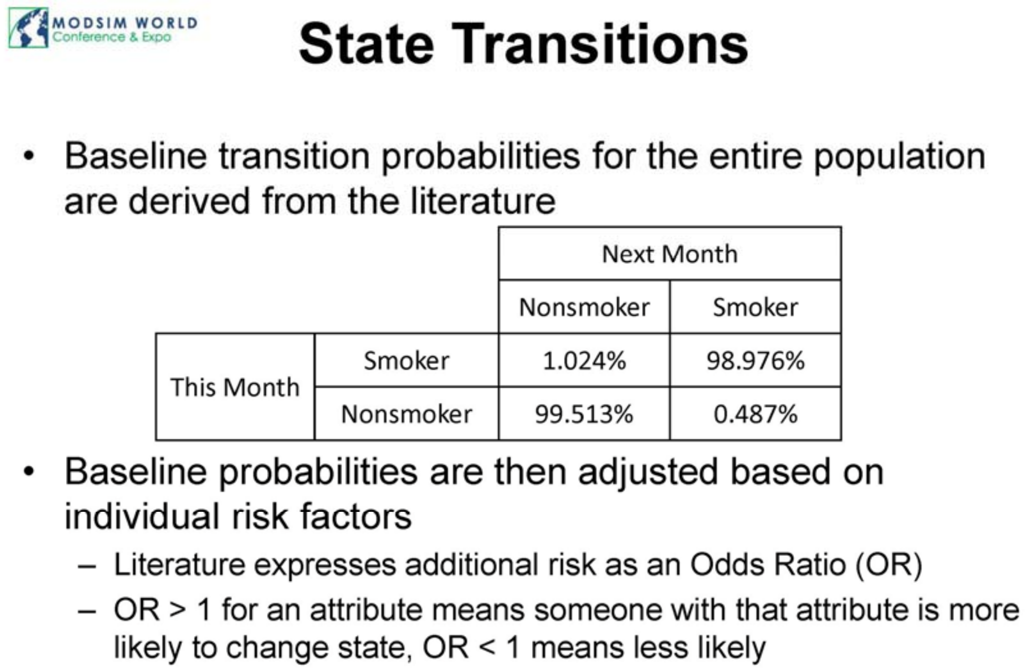
Individuals and States

- Individuals in the simulation have several attributes that describe their state at any given time
  - Smoker or nonsmoker
  - Age
  - Gender
  - Months smoked (total and consecutive)
- Individuals also retain social relationships which affect smoking behavior
  - Parent (single parent, smoking status recorded)
  - Peers (links to “nearby” individuals close to age)



Time Ticks

- Each month (a “tick” of the simulation clock) an individual’s state is updated
  - Age and other tracking variables are incremented
  - Smoking is commenced or ceased based on probabilities
  - In the extended model, an individual may develop disease based on probabilities
- Probabilities of changing states are affected by attributes of the individual and their social relationships
  - Parent and peer smoking status affects behavior
  - Age, Gender, prior smoking status has impact on risk

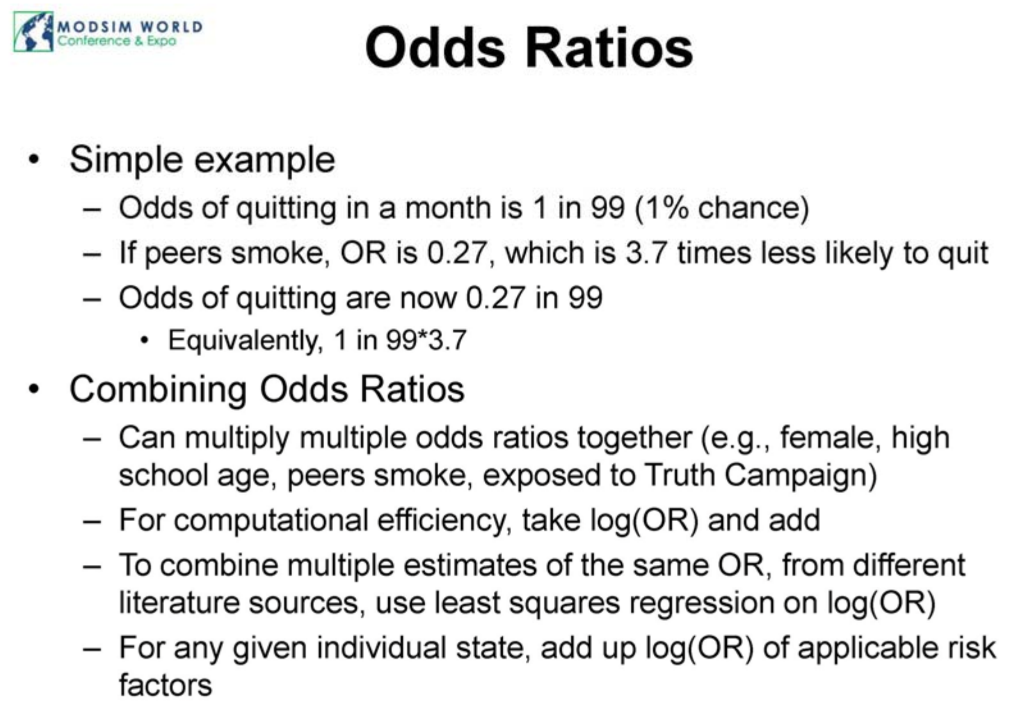


State Transitions

- Baseline transition probabilities for the entire population are derived from the literature

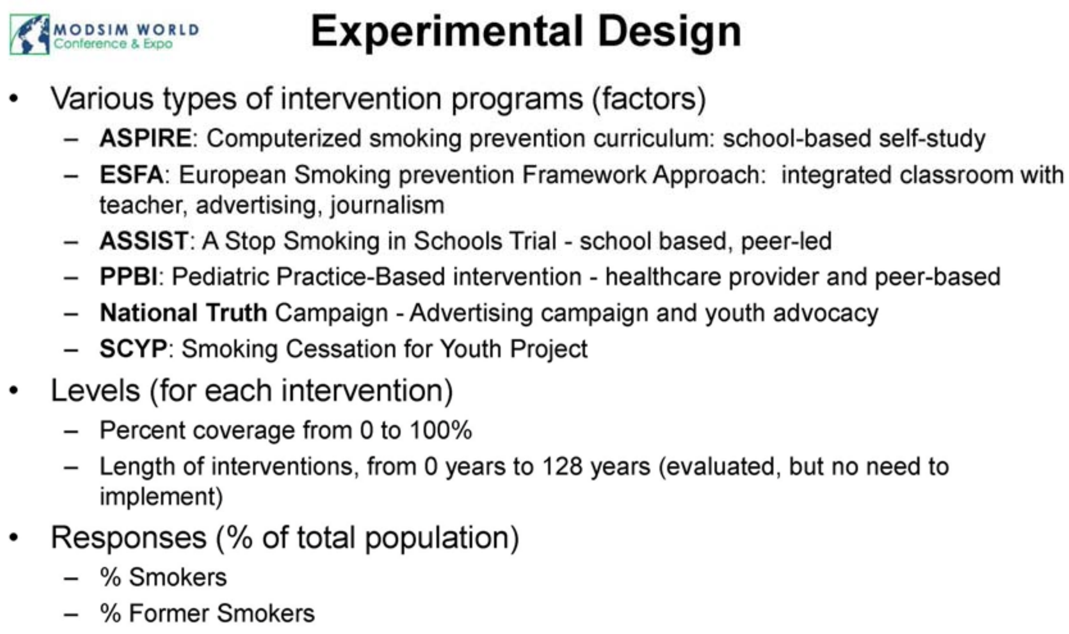
		Next Month	
		Nonsmoker	Smoker
This Month	Smoker	1.024%	98.976%
	Nonsmoker	99.513%	0.487%

- Baseline probabilities are then adjusted based on individual risk factors
  - Literature expresses additional risk as an Odds Ratio (OR)
  - OR > 1 for an attribute means someone with that attribute is more likely to change state, OR < 1 means less likely



Odds Ratios

- Simple example
  - Odds of quitting in a month is 1 in 99 (1% chance)
  - If peers smoke, OR is 0.27, which is 3.7 times less likely to quit
  - Odds of quitting are now 0.27 in 99
    - Equivalently, 1 in 99*3.7
- Combining Odds Ratios
  - Can multiply multiple odds ratios together (e.g., female, high school age, peers smoke, exposed to Truth Campaign)
  - For computational efficiency, take log(OR) and add
  - To combine multiple estimates of the same OR, from different literature sources, use least squares regression on log(OR)
  - For any given individual state, add up log(OR) of applicable risk factors



Experimental Design

- Various types of intervention programs (factors)
  - **ASPIRE**: Computerized smoking prevention curriculum: school-based self-study
  - **ESFA**: European Smoking prevention Framework Approach: integrated classroom with teacher, advertising, journalism
  - **ASSIST**: A Stop Smoking in Schools Trial - school based, peer-led
  - **PPBI**: Pediatric Practice-Based intervention - healthcare provider and peer-based
  - **National Truth** Campaign - Advertising campaign and youth advocacy
  - **SCYP**: Smoking Cessation for Youth Project
- Levels (for each intervention)
  - Percent coverage from 0 to 100%
  - Length of interventions, from 0 years to 128 years (evaluated, but no need to implement)
- Responses (% of total population)
  - % Smokers
  - % Former Smokers

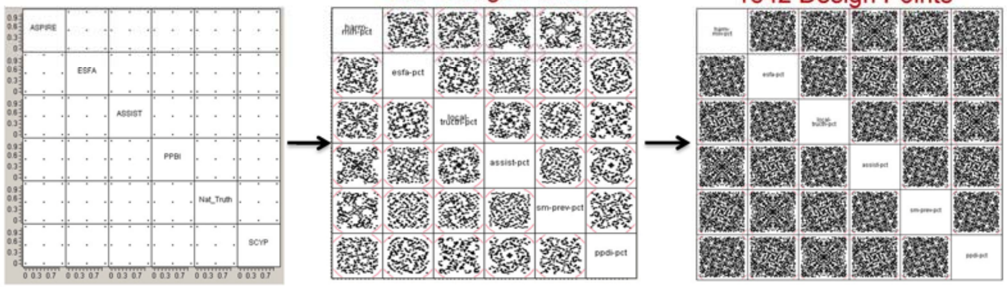
MODSIM WORLD Conference & Expo

Evolution of Design

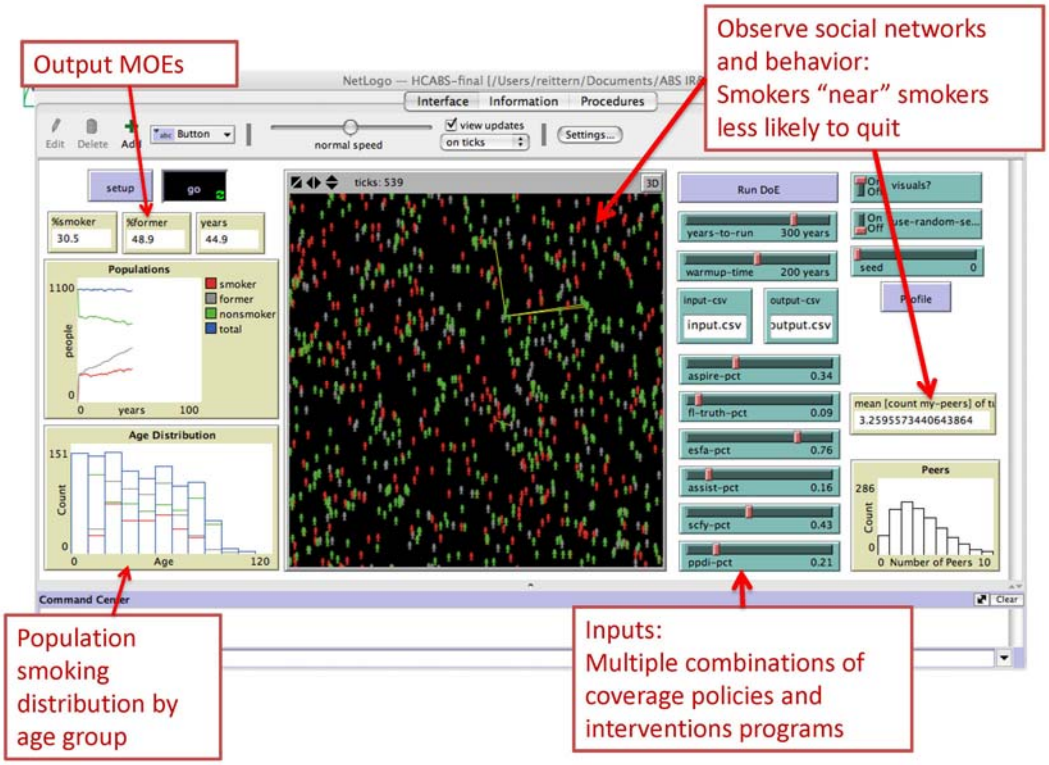
3⁶ Factorial Design
729 design points; 12 hour runtime

1 x 257 NOLH Design
257 Design Points

6 x 257 NOLH Design
(Rotated)
1542 Design Points



Roughly twice as many points as 3⁶ factorial with huge design space coverage



Output MOEs

%smoker 30.5

%former 48.9

years 44.9

years-to-run 300 years

warmup-time 200 years

input-csv input.csv

output-csv output.csv

aspire-pct 0.34

fl-truth-pct 0.09

esfa-pct 0.76

assist-pct 0.16

scfy-pct 0.43

ppdi-pct 0.21

seed 0

mean [count my-peers] of t
3.2595573440643864

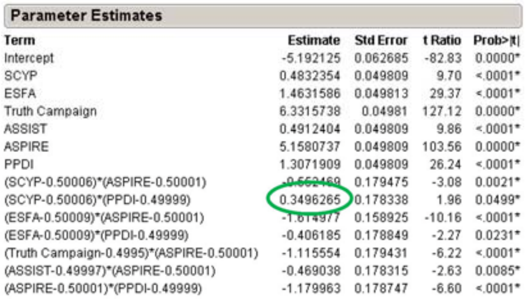
Observe social networks and behavior:
Smokers "near" smokers less likely to quit

Population smoking distribution by age group

Inputs:
Multiple combinations of coverage policies and interventions programs

Initial Analysis Results

- Multivariate Regression analysis
  - All 6 interventions as dependent variables, with all 2-way interactions
  - Decrease in % smokers as independent variable (positive is good)
  - Expected results:
    - positive coefficients for each intervention
    - Negative coefficients for interactions due to diminishing returns
  - Actual results:
    - SCYP * PPDI positive interaction
    - Using both together better than each one separately

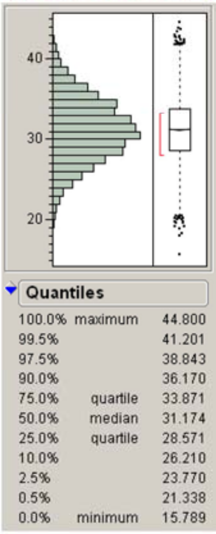


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-5.192125	0.062685	-82.83	0.0000*
SCYP	0.4832354	0.049809	9.70	<.0001*
ESFA	1.4631586	0.049813	29.37	<.0001*
Truth Campaign	6.3315738	0.04981	127.12	0.0000*
ASSIST	0.4912404	0.049809	9.86	<.0001*
ASPIRE	5.1580737	0.049809	103.56	0.0000*
PPDI	1.3071909	0.049809	26.24	<.0001*
(SCYP-0.50006)*(ASPIRE-0.50001)	-0.552469	0.179475	-3.08	0.0021*
(SCYP-0.50006)*(PPDI-0.49999)	0.3496265	0.178338	1.96	0.0499*
(ESFA-0.50009)*(ASPIRE-0.50001)	-1.614977	0.158925	-10.16	<.0001*
(ESFA-0.50009)*(PPDI-0.49999)	-0.406185	0.178849	-2.27	0.0231*
(Truth Campaign-0.4995)*(ASPIRE-0.50001)	-1.115554	0.179431	-6.22	<.0001*
(ASSIST-0.49997)*(ASPIRE-0.50001)	-0.469038	0.178315	-2.63	0.0085*
(ASPIRE-0.50001)*(PPDI-0.49999)	-1.179963	0.178747	-6.60	<.0001*

Sampling of Model Response

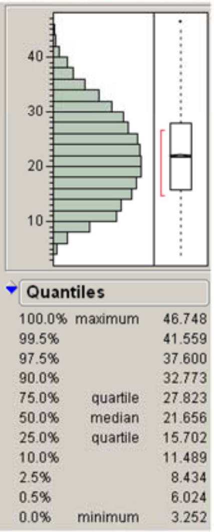
% Smokers (pre-interventions)



Quantiles

100.0%	maximum	44.800
99.5%		41.201
97.5%		38.843
90.0%		36.170
75.0%	quartile	33.871
50.0%	median	31.174
25.0%	quartile	28.571
10.0%		26.210
2.5%		23.770
0.5%		21.338
0.0%	minimum	15.789

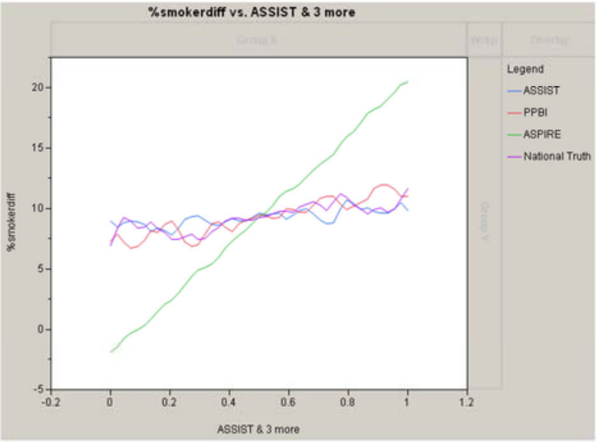
% Smokers (post interventions)



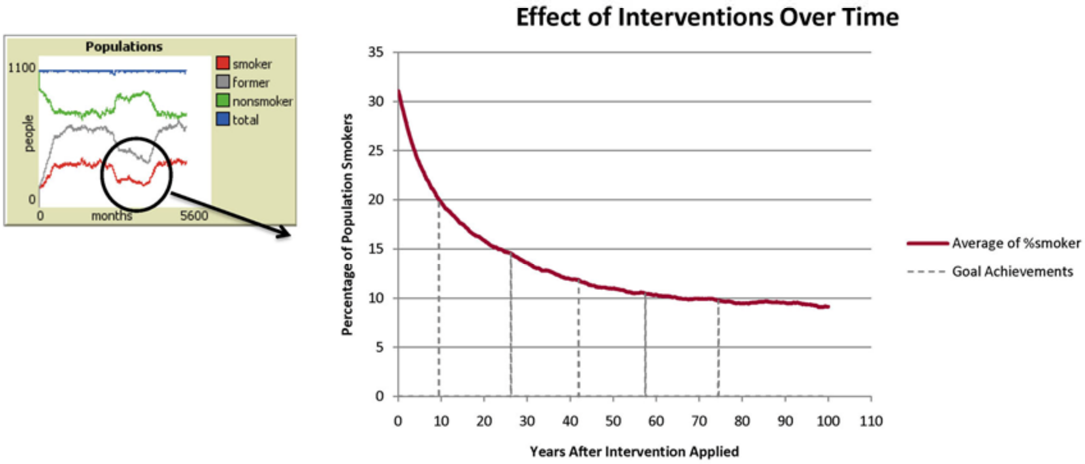
Quantiles

100.0%	maximum	46.748
99.5%		41.559
97.5%		37.600
90.0%		32.773
75.0%	quartile	27.823
50.0%	median	21.656
25.0%	quartile	15.702
10.0%		11.489
2.5%		8.434
0.5%		6.024
0.0%	minimum	3.252

Effects of Interventions



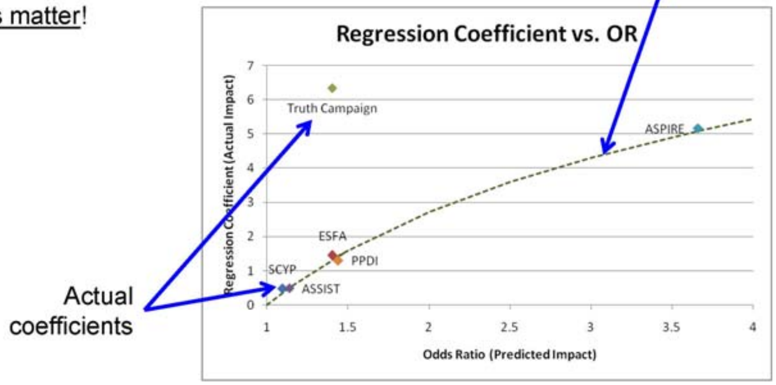
(Zooming in on timeframe when interventions took effect)



Impact of interactions on predictions

- Tested up to 6-way interactions
  - Statistically significant interactions up to 5th level
  - You can't just predict response from the OR
    - Actual response impacted by interactions
    - Risk factors matter!

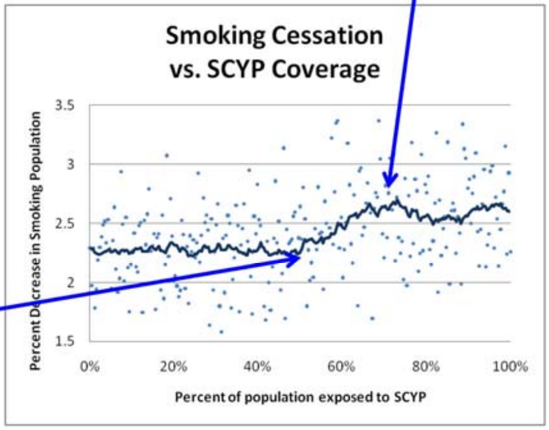
Predicted coefficients



Actual coefficients

A closer look at SCYP

- SCYP shows a clear "threshold effect"
  - PPDI Interaction highlighted this sensitivity to other interventions
  - Minimum and maximum effective level
  - Dependent on which other interventions are employed

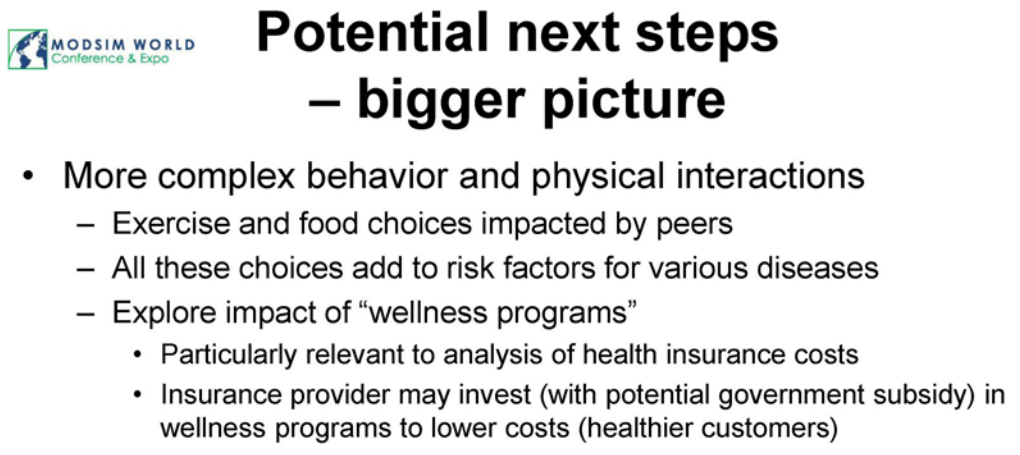


Invest no more than this

Invest no less than this

Potential next steps – just for smoking

- Simulation results used to populate "response surface"
  - Lots of threshold effects for other interventions at various combinations
    - 7-dimensional, so we can't show you here
  - Given costs of each intervention, along with cost constraints, can use optimization methods to find best mix at each investment level
  - Pareto frontier of optimal intervention mixes can inform decisions on overall investment level
- Additional simulation exploration of "non-overlapping" multiple interventions
  - Each individual might only experience one intervention, but peers may experience others
  - Potential to mitigate negative interactions due to "over-intervening"



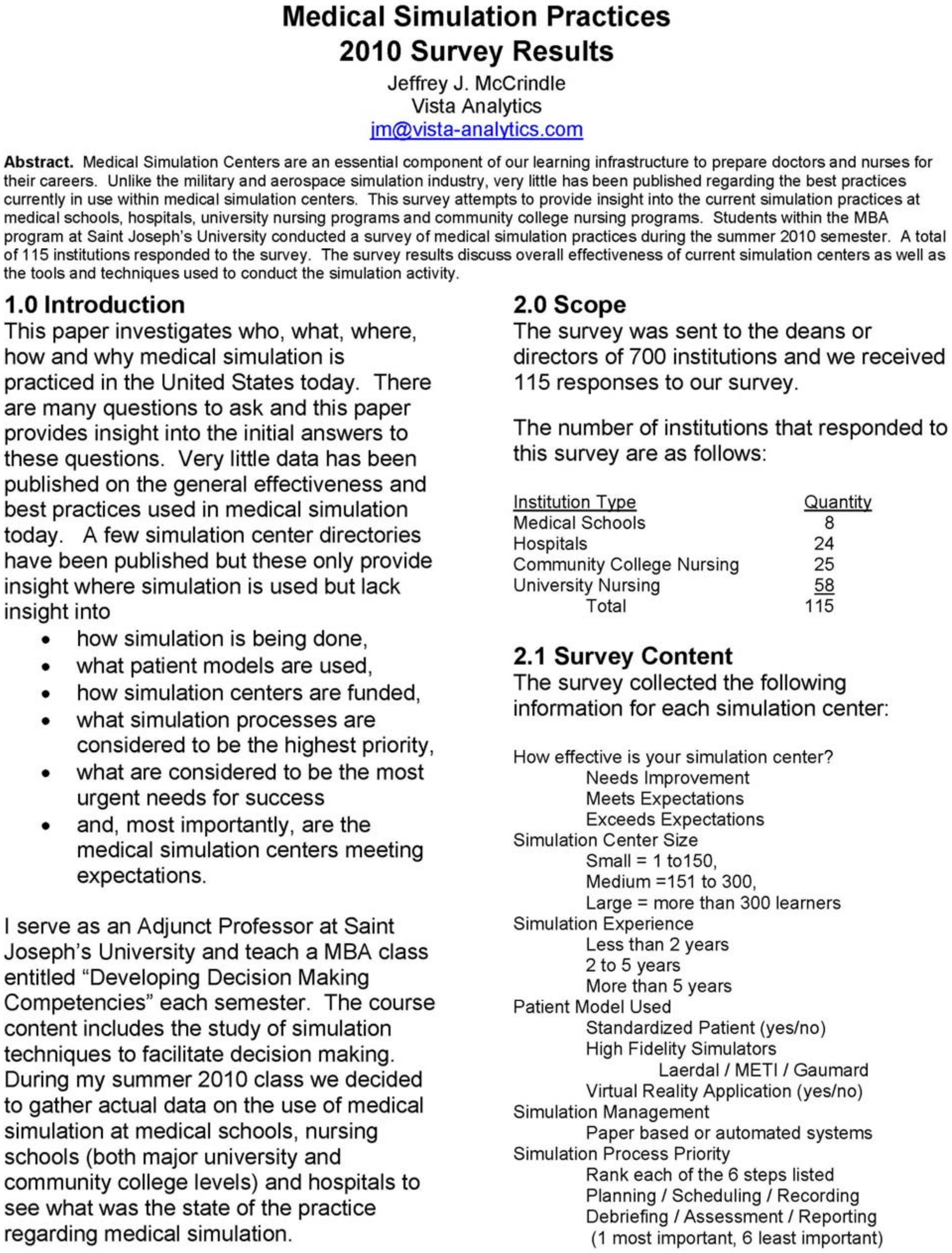
Potential next steps – bigger picture

- More complex behavior and physical interactions
  - Exercise and food choices impacted by peers
  - All these choices add to risk factors for various diseases
  - Explore impact of “wellness programs”
    - Particularly relevant to analysis of health insurance costs
    - Insurance provider may invest (with potential government subsidy) in wellness programs to lower costs (healthier customers)

Questions?

3.0 HEALTH & MEDICINE TRACK

3.1 Medical Simulation Practices 2010 Survey Results



Medical Simulation Practices 2010 Survey Results

Jeffrey J. McCrindle
Vista Analytics
jm@vista-analytics.com

Abstract. Medical Simulation Centers are an essential component of our learning infrastructure to prepare doctors and nurses for their careers. Unlike the military and aerospace simulation industry, very little has been published regarding the best practices currently in use within medical simulation centers. This survey attempts to provide insight into the current simulation practices at medical schools, hospitals, university nursing programs and community college nursing programs. Students within the MBA program at Saint Joseph's University conducted a survey of medical simulation practices during the summer 2010 semester. A total of 115 institutions responded to the survey. The survey results discuss overall effectiveness of current simulation centers as well as the tools and techniques used to conduct the simulation activity.

1.0 Introduction

This paper investigates who, what, where, how and why medical simulation is practiced in the United States today. There are many questions to ask and this paper provides insight into the initial answers to these questions. Very little data has been published on the general effectiveness and best practices used in medical simulation today. A few simulation center directories have been published but these only provide insight where simulation is used but lack insight into

- how simulation is being done,
- what patient models are used,
- how simulation centers are funded,
- what simulation processes are considered to be the highest priority,
- what are considered to be the most urgent needs for success
- and, most importantly, are the medical simulation centers meeting expectations.

I serve as an Adjunct Professor at Saint Joseph's University and teach a MBA class entitled "Developing Decision Making Competencies" each semester. The course content includes the study of simulation techniques to facilitate decision making. During my summer 2010 class we decided to gather actual data on the use of medical simulation at medical schools, nursing schools (both major university and community college levels) and hospitals to see what was the state of the practice regarding medical simulation.

2.0 Scope

The survey was sent to the deans or directors of 700 institutions and we received 115 responses to our survey.

The number of institutions that responded to this survey are as follows:

Institution Type	Quantity
Medical Schools	8
Hospitals	24
Community College Nursing	25
University Nursing	58
Total	115

2.1 Survey Content

The survey collected the following information for each simulation center:

How effective is your simulation center?
Needs Improvement
Meets Expectations
Exceeds Expectations
Simulation Center Size
Small = 1 to150,
Medium =151 to 300,
Large = more than 300 learners
Simulation Experience
Less than 2 years
2 to 5 years
More than 5 years
Patient Model Used
Standardized Patient (yes/no)
High Fidelity Simulators
Laerdal / METI / Gaumard
Virtual Reality Application (yes/no)
Simulation Management
Paper based or automated systems
Simulation Process Priority
Rank each of the 6 steps listed
Planning / Scheduling / Recording
Debriefing / Assessment / Reporting
(1 most important, 6 least important)

Select which is the most urgent need
Standardized Scenario Content
Return on Investment Case Studies
Lower Cost Simulation Solutions
What is the annual cost of running your simulation center?
$0 to $50,000
$50,001 to $200,000
$200,001 to $400,000
Over $400,000
How do you fund your simulation center?
Grants (yes/no)
Strategic Donors (yes/no)
Operational Budget (yes/no)
Do you share your simulation center with others? (yes / no / interested)
Do you include team training in your simulation scenarios? (yes / no / interested)
Do you model Electronic Health Records in your simulated scenarios? (yes/no)
Do you use video of recorded simulations during debriefing? (yes/no)

3.0 Survey Results

3.1 Simulation Center Effectiveness

The survey results show that the majority of simulation centers meet or exceed management expectations. Two areas that require further investigation are community college nursing programs where 35% of responses indicate that their simulation center needs improvement and medical schools where 62% indicate improvement is needed. Given the small sample size for medical schools (8) the high percentage can be misleading – more medical school data should be collected to explore this result.

EE = Exceeds Expectations
ME = Meets Expectations
NI = Needs Improvement

Sim Center Effectiveness	EE	ME	NI
Medical Schools	13%	25%	62%
Hospitals	36%	41%	23%
Community College Nursing	30%	35%	35%
University Nursing	25%	53%	22%

Fig 3.1 Simulation Center Effectiveness

3.2 Simulation Center Experience

Medical simulation has been actively used at most medical schools for over 5 years. Over 78% of university nursing programs have had established simulation centers in place for 2 or more years. Hospitals show an even distribution over the three experience levels. Community college nursing program results show that over 82% of these simulation centers have been in place for 5 years or less.

<2 = less than 2 years
2 – 5 = between 2 and 5 years
5+ = more than 5 years

Sim Center Experience	<2	2 - 5	5+
Medical Schools	13%	25%	62%
Hospitals	32%	32%	37%
Community College Nursing	30%	52%	17%
University Nursing	22%	40%	38%

Fig 3.2 Simulation Center Experience

3.3 Sim Center Size (# Learners)

Survey results show that hospitals typically support more learners than medical schools, university nursing and community college nursing programs. More than 70% of hospitals, university nursing and community college nursing centers had more than 150 learners.

As we started receiving the survey results we realized that our range of possible values did not account for very large simulation centers. A few centers commented that they had significantly more than 300 learners.

S = less than 150 learners
M = between 150 and 300 learners
L = more than 300 learners

Sim Center Size	S	M	L
Medical Schools	0%	63%	37%
Hospitals	21%	21%	58%
Community College Nursing	30%	48%	22%
University Nursing	24%	42%	34%

Fig 3.3 Number of Learners

3.4 Simulation Process Priority

Medical schools, hospitals, university nursing and community college nursing programs all demonstrated the same the survey results when asked what simulation processes were the most important. Clearly planning and debriefing processes are considered to be the highest priority.

Recording simulation activity on video and reporting were not considered as important which was a little surprising when considering these components constitute evidence of competency and provide support for longitudinal studies.

Ranking 1 = highest priority to 6 = lowest priority

Simulation Process	Average Ranking
Planning	1.8
Scheduling	3.3
Recording	4.3
Debriefing	2.1
Assessment	3.8
Reporting	4.7

Fig 3.4 Simulation Process Priority

3.5 Patient Models Used

The survey also asked each institution what patient models were used. It was interesting to see that almost all institutions used a combination of standardized patient, high fidelity simulators, and/or virtual reality training aids. Clearly, high fidelity simulators are used in almost all centers. We expected to see higher results for the use of Standardized Patients in community college nursing programs. The lower results may be associated with the relatively high cost of managing a Standardized Patient program.

SP = standardized patients
HFS = Laerdal, METI and/or Gaumard simulator
VR = virtual reality application

Sim Center	SP	HFS	VR
Medical Schools	75%	88%	50%
Hospitals	46%	96%	38%
Community College Nursing	26%	96%	22%
University Nursing	47%	96%	40%

Fig 3.5 Patient Models by Sim Center Type

3.6 High Fidelity Simulator Use

For those institutions using high fidelity simulators we asked specifically what vendors supplied the simulators that were used in their center. Laerdal clearly is the market share leader across the surveyed institutions. With the exception of medical schools, Gaumard appears to have secured the second market share position. It will be interesting to watch these market share statistics over time as each vendor introduces new simulator models.

L = Laerdal M = METI G = Gaumard

Sim Center	L	M	G
Medical Schools	86%	57%	29%
Hospitals	78%	43%	52%
Community College Nursing	96%	21%	29%
University Nursing	89%	29%	58%

Fig 3.6 High Fidelity Simulators

3.7 Mix of High Fidelity Simulators

Very few simulation centers use simulators from just one vendor. The following paragraphs detail how often each vendor's simulator products were used at the same institution along with other vendor's products.

Medical Schools	
Laerdal only	43%
Laerdal and METI	14%
Laerdal and Gaumard	0%
Laerdal, METI and Gaumard	29%
METI only	14%
METI and Gaumard	0%
Gaumard only	0%

Hospitals	
Laerdal only	26%
Laerdal and METI	4%
Laerdal and Gaumard	31%
Laerdal, METI and Gaumard	17%
METI only	17%
METI and Gaumard	5%
Gaumard only	0%

Community College Nursing	
Laerdal only	54%
Laerdal and METI	17%
Laerdal and Gaumard	21%
Laerdal, METI and Gaumard	4%
METI only	0%
METI and Gaumard	0%
Gaumard only	4%

University Nursing	
Laerdal only	29%
Laerdal and METI	9%
Laerdal and Gaumard	40%
Laerdal, METI and Gaumard	11%
METI only	4%
METI and Gaumard	5%
Gaumard only	2%

3.8 Simulation Management

Institutions were asked whether they used a paper based management approach for their simulation center or did they use automated simulation management solution (either a collection of products or an integrated systems). Surprisingly, more than half of the community college nursing programs surveyed are still using paper based management approaches.

Considering the effort that is associated with planning, scheduling, assessing student performance and reporting, the lack of automation may be a limiting factor in the effective use of medical simulation at community colleges.

Sim Center	Paper based	Automated
Medical Schools	50%	50%
Hospitals	29%	71%
Community College Nursing	54%	46%
University Nursing	24%	76%

Fig 3.8 Simulation Management

3.9 Most Urgent Need

The survey asked institutions to select which of three items was the most urgent need for growth in use of medical simulation. The three options included (1) the need for standardized simulation scenario content, (2) proven return on investment (ROI) studies, and (3) lower cost for simulation technology. Hospitals and community college nursing programs identified standardized simulation scenario content as the most urgent need where university nursing programs seek proven ROI studies so they can justify the expansion of their simulation centers.

C = Standardized scenario content
R = Proven ROI Case Studies
L = Lower cost simulation solutions

Sim Center	C	R	L
Medical Schools	0%	62%	38%
Hospitals	46%	37%	17%
Community College Nursing	48%	13%	39%
University Nursing	33%	43%	24%

Fig 3.9 Most Urgent Need by Institution Type

3.10 EMR Simulation

Electronic Medical Records and Electronic Health Records systems will be an important element for learners to practice with as part of their simulation experience.

Most institutions want to practice in a generic EMR system environment so their learners can effectively operate with any commercial EMR implementation.

Based on the survey results it is clear that the majority of simulation centers recognize the need for EMR training.

Yes = plan to simulate EMR systems in 2010
No = No plans to simulate EMR use

Sim Center	Yes	No
Medical Schools	63%	37%
Hospitals	71%	29%
Community College Nursing	70%	30%
University Nursing	92%	8%

Fig 3.10 EMR Simulation by Institution Type

3.12 Sharing your Simulation Center

Creating and operating a simulation center requires a significant investment. We have observed a growing trend for institutions to share their simulation centers with outside users. This survey asked each institution whether they were actively sharing their center, had no interest in sharing or were not currently sharing their center but had interest in exploring this option.

Medical schools and hospitals appear to be actively engaged in sharing their simulation resources. Both university and community college nursing programs have significant interest in exploring the benefits of sharing their simulation center resources.

With automated simulation management systems that provide accounting records for chargeback and separate tracking of 3rd party simulation planning, scheduling, assessment and reporting, sharing a simulation center is now easily accomplished.

Yes = actively sharing
No = no sharing
I = not currently sharing but interested

Sim Center	Yes	No	I
Medical Schools	63%	12%	25%
Hospitals	54%	8%	38%
Community College Nursing	13%	26%	61%
University Nursing	34%	21%	45%

Fig 3.12 Sharing Sim Center Resources

3.13 Team Training

The ability to clearly and concisely communicate is as important as the technical skills learned in medical schools and nursing schools. As the survey results show, the majority of simulation centers understand the importance of team communication and have, or intend to have, team training incorporated into their programs.

There are many approaches to team training. The TEAMSTEPPS framework (http://teamsstepps.ahrq.gov) is an example of the team training programs that are being implemented in both government and commercial institutions.

Community college nursing programs are slightly behind in the adoption of team training practices but even in this case 78% of these institutions are focused on team training.

Yes = actively implements team training
No = no team training
I = not currently doing team training but interested

Sim Center	Yes	No	I
Medical Schools	63%	12%	25%
Hospitals	84%	4%	12%
Community College Nursing	61%	22%	17%
University Nursing	67%	7%	26%

Fig 3.13 Team Training

3.14 Use of Video during Debriefing

As we observed in the simulation process priority discussion in section 3.4, debriefing is a very high priority component to medical simulation.

Almost every center that I have met has identified debriefing to be the richest learning experience for the learner. Given that debriefing is very important it is interesting to see different approaches regarding how debriefing is conducted.

Until recently, video of simulation activity was captured on VHS tapes and use of video for debriefing was not very compelling. With the introduction of digital audio video systems that can provide rapid access to video and the ability to bookmark key time tagged events the use of video to effectively support debriefing is now possible.

Survey results that show 30 to 40% of the institutions are not using video for debriefing. The question remains whether this is due to an older video system or a process preference to debrief simulation activity without video.

Yes = using video during debriefing
No = not using video

Sim Center	Yes	No
Medical Schools	62%	38%
Hospitals	67%	33%
Community College Nursing	70%	30%
University Nursing	68%	32%

Fig 3.14 Use of Video for Debriefing

3.15 Funding Sim Center Operations

The survey asked institutions what their annual cost was for operating their simulation center. Upon further review the survey should have clearly broken out the following costs

- to initially implement the center,
- the ongoing cost to support the simulator technology,
- the ongoing cost to cover standardized patients,
- and the cost for internal simulation center staff.

As currently implemented, the survey shows that medical schools report the highest annual cost for operating their simulation center, followed by university nursing programs, community college nursing programs and hospitals.

Sim Center	< 50K	50K to 200K	200K to 400K	Over 400K
Medical Schools	12%	25%	25%	38%
Hospitals	49%	23%	5%	23%
Community College Nursing	52%	30%	13%	5%
University Nursing	33%	49%	12%	6%

Fig 3.15 Annual Sim Center Funding

3.16 Funding Sources

We asked institutions if they used strategic donors, grants, and/or use their operational budgets to fund their simulation centers. Fig 3.16 shows that strategic donors are used more frequently at university nursing programs and hospitals. Grants are used heavily by all institution types except medical schools. Operating budget funding for simulation is used heavily by all institutions. It would have been interesting to see how institutions would have described the percent of funds that were obtained from each of these funding sources.

Sim Center	Strategic Donor	Grants	Op Budget
Medical Schools	14%	29%	100%
Hospitals	45%	73%	77%
Community College Nursing	30%	70%	61%
University Nursing	50%	66%	72%

Fig 3.16 Funding Sources

4.0 Conclusion

Medical simulation offers significant value to institutions as they prepare learners for their careers. Simulation is a well defined discipline in the military and aerospace industries where standards and best practices have been established over many years. Medical simulation tools and techniques are relatively new but great progress has been made in a relatively short time.

Standards organizations are actively work to bring medical simulation institutions together to share best practices. This survey was an initial attempt to capture how medical simulation is being practiced in 2010. As we received survey responses from institutions we realized that we would have asked questions in a slightly different way and asked more questions to gain additional insight into how and why an institution conducts simulation as they do.

This paper only discussed the high level results of this medical simulation survey. More detailed data analysis – for example how simulation center effectiveness varied by size of institution, types of patient models used, simulation management approaches - is currently underway. The results will be posted at www.vista-analytics.com in mid October 2010.

5.0 Acknowledgements

The Medical Simulation Survey was conducted by the MBA students at Saint Joseph's University. Their professionalism and persistence is greatly appreciated. The research team included:

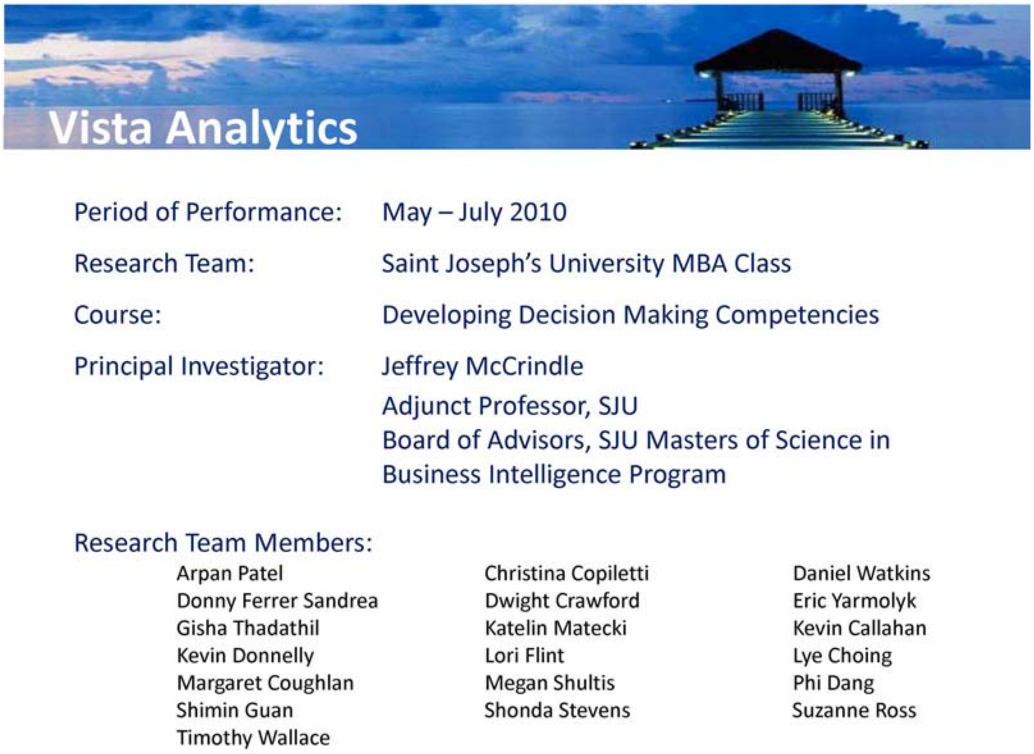
Arpan Patel
Christina Copiletti
Daniel Watkins
Donny Ferrer Sandrea
Dwight Crawford
Eric Yarmolyk
Gisha Thadathil
Katelin Matecki
Kevin Callahan
Kevin Donnelly
Lori Flint
Lye Choing
Margaret Coughlan
Megan Shultis
Phi Dang
Shimin Guan
Shonda Stevens
Suzanne Ross
Timothy Wallace

Vista Analytics
Medical Simulation Practices
2010 Survey

Vista Analytics

Medical Simulation Practices - 2010 Survey

To improve the understanding of current medical simulation practice at medical schools, hospitals and nursing schools.



Vista Analytics

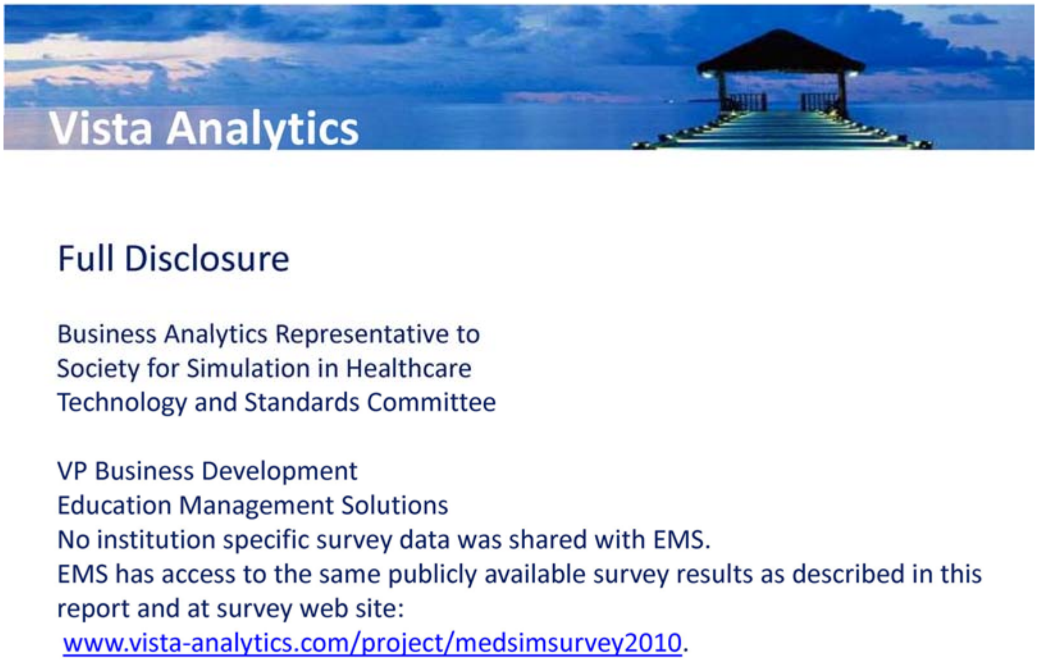
Period of Performance:	May – July 2010
Research Team:	Saint Joseph's University MBA Class
Course:	Developing Decision Making Competencies
Principal Investigator:	Jeffrey McCrindle Adjunct Professor, SJU Board of Advisors, SJU Masters of Science in Business Intelligence Program

Research Team Members:

Arpan Patel
Donny Ferrer Sandrea
Gisha Thadathil
Kevin Donnelly
Margaret Coughlan
Shimin Guan
Timothy Wallace

Christina Copiletti
Dwight Crawford
Katelin Matecki
Lori Flint
Megan Shultis
Shonda Stevens

Daniel Watkins
Eric Yarmolyk
Kevin Callahan
Lye Choing
Phi Dang
Suzanne Ross

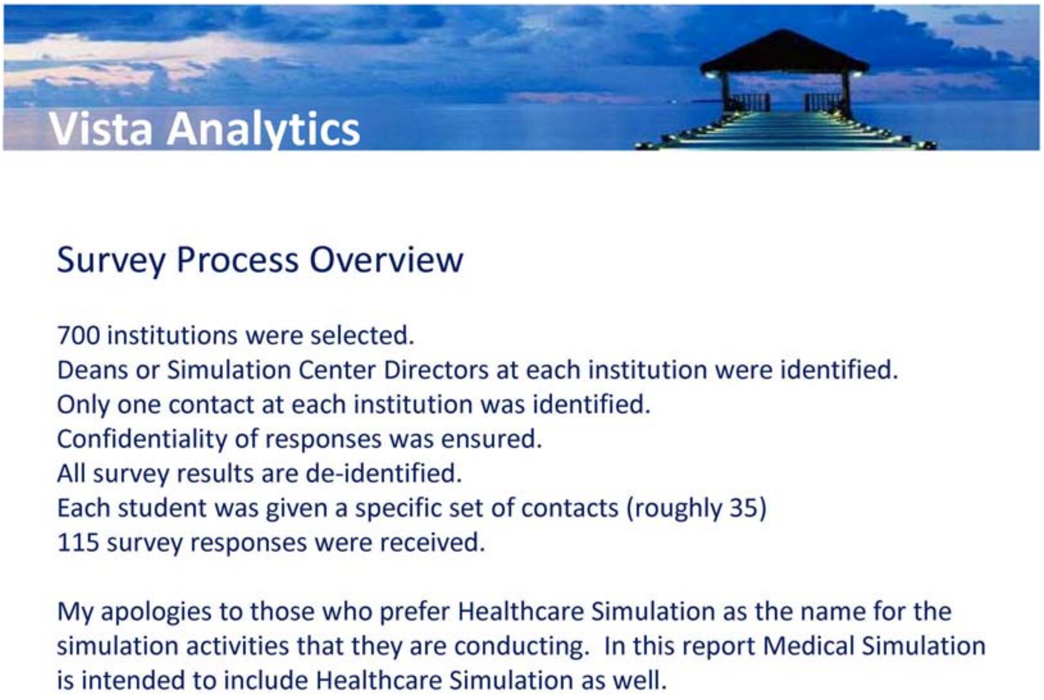


Vista Analytics

Full Disclosure

Business Analytics Representative to
Society for Simulation in Healthcare
Technology and Standards Committee

VP Business Development
Education Management Solutions
No institution specific survey data was shared with EMS.
EMS has access to the same publicly available survey results as described in this report and at survey web site:
www.vista-analytics.com/project/medsimsurvey2010.

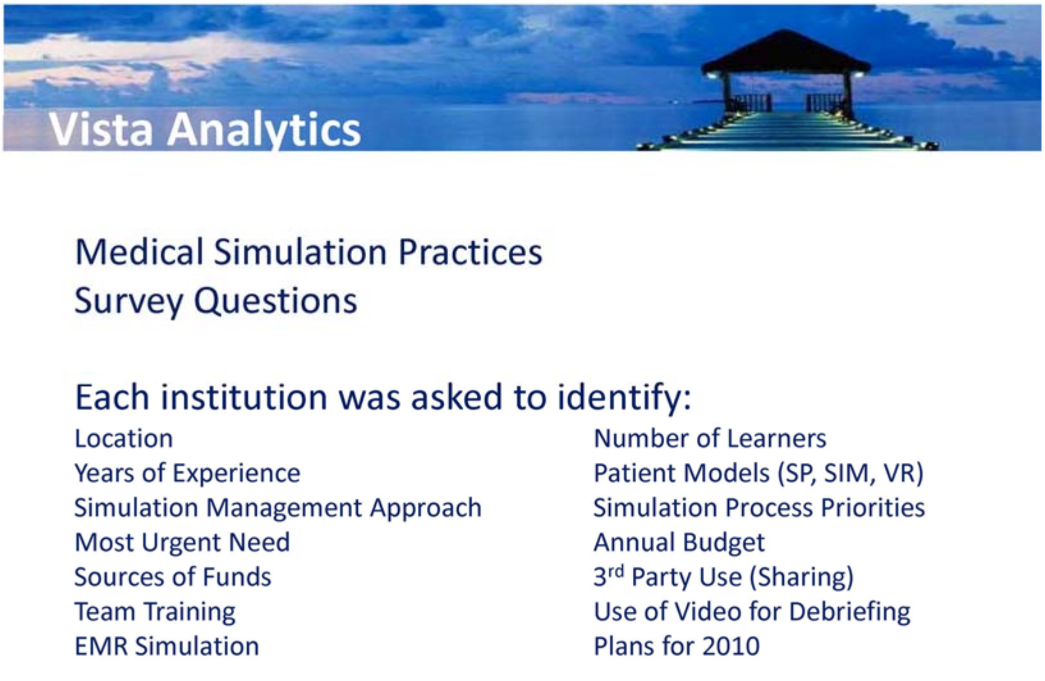


Vista Analytics

Survey Process Overview

700 institutions were selected.
Deans or Simulation Center Directors at each institution were identified.
Only one contact at each institution was identified.
Confidentiality of responses was ensured.
All survey results are de-identified.
Each student was given a specific set of contacts (roughly 35)
115 survey responses were received.

My apologies to those who prefer Healthcare Simulation as the name for the simulation activities that they are conducting. In this report Medical Simulation is intended to include Healthcare Simulation as well.



Vista Analytics

Medical Simulation Practices
Survey Questions

Each institution was asked to identify:

Location
Years of Experience
Simulation Management Approach
Most Urgent Need
Sources of Funds
Team Training
EMR Simulation

Number of Learners
Patient Models (SP, SIM, VR)
Simulation Process Priorities
Annual Budget
3rd Party Use (Sharing)
Use of Video for Debriefing
Plans for 2010

Vista Analytics

Institutions that responded included:

8	Medical Schools
58	University Nursing Programs
25	Community College Nursing Programs
24	Hospitals

115	Total Institutions

Vista Analytics

Institutions rated the effectiveness of their simulation centers:

Exceeds expectations (EE)
Meets expectations (ME)
Needs Improvement (NI)

Simulation Center Effectiveness	EE	ME	NI
Medical Schools	13%	25%	62%
Hospitals	36%	41%	23%
Community College Nursing	30%	35%	35%
University Nursing	25%	53%	22%

How many years have you conducted medical simulation at your institution?

0 to 2 years
2 to 5 years
More than 5 years

Simulation Center Experience	<2	2 - 5	5+
Medical Schools	13%	25%	62%
Hospitals	32%	32%	37%
Community College Nursing	30%	52%	17%
University Nursing	22%	40%	38%

Rate the priority of each of the following simulation processes (1 = highest, 6 = lowest)

	Mean	Std Dev
Planning	1.8	1.1
Scheduling	3.3	1.6
Recording	4.3	1.5
Debriefing	2.1	1.3
Assessment	3.8	1.4
Reporting	4.7	1.3

What Patient Models Do You Use?

Standardized Patients (SP)
High Fidelity Simulators (SIM)
Virtual Reality Applications (VR)

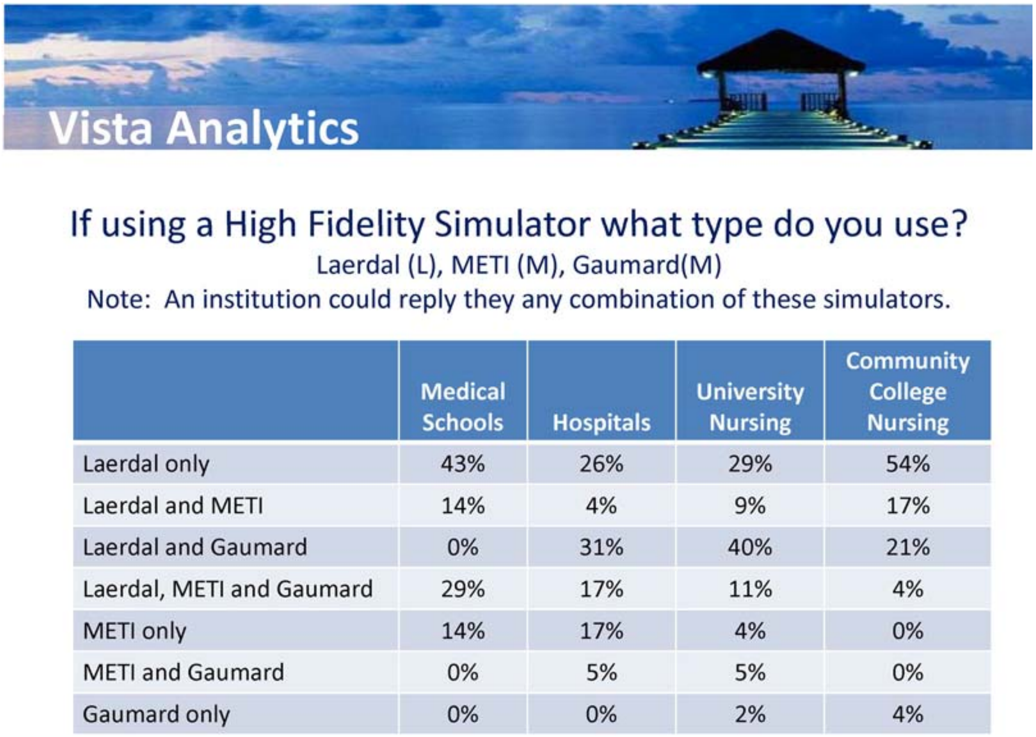
Note: An institution could reply they any combination of these patient models

Simulation Center	SP	SIM	VR
Medical Schools	75%	88%	50%
Hospitals	46%	96%	38%
Community College Nursing	26%	96%	22%
University Nursing	47%	96%	40%

If using a High Fidelity Simulator what type do you use?

Laerdal (L), METI (M), Gaumard(M)
Note: An institution could reply they any combination of these simulators.

	Medical Schools	Hospitals	University Nursing	Community College Nursing
Laerdal	86%	78%	89%	96%
METI	57%	39%	24%	21%
Gaumard	29%	48%	53%	29%



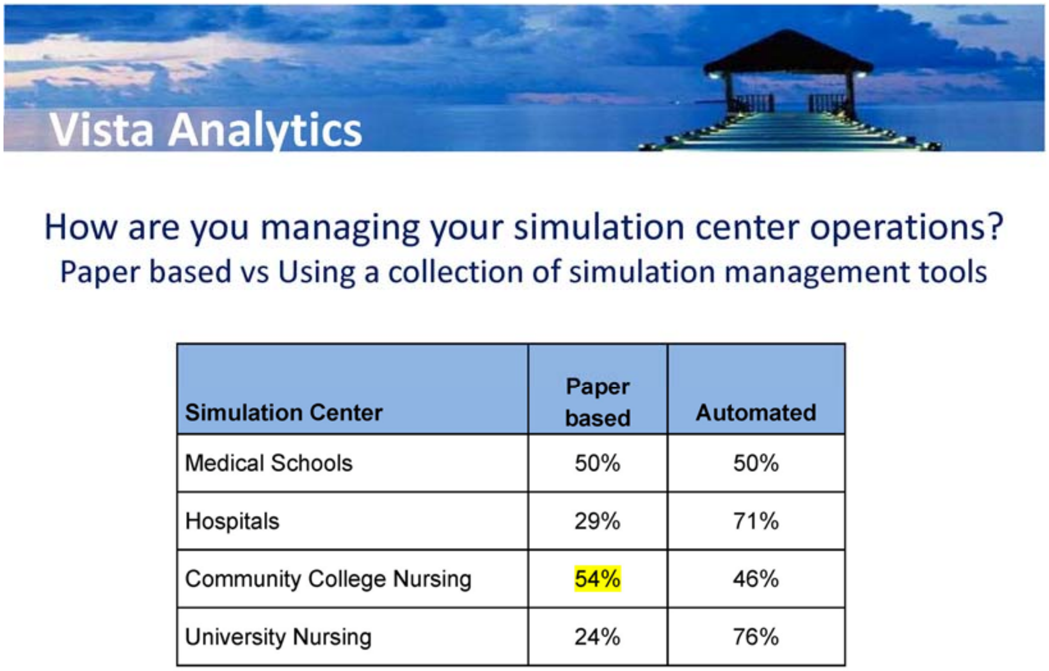
Vista Analytics

If using a High Fidelity Simulator what type do you use?

Laerdal (L), METI (M), Gaumard(M)

Note: An institution could reply they any combination of these simulators.

	Medical Schools	Hospitals	University Nursing	Community College Nursing
Laerdal only	43%	26%	29%	54%
Laerdal and METI	14%	4%	9%	17%
Laerdal and Gaumard	0%	31%	40%	21%
Laerdal, METI and Gaumard	29%	17%	11%	4%
METI only	14%	17%	4%	0%
METI and Gaumard	0%	5%	5%	0%
Gaumard only	0%	0%	2%	4%

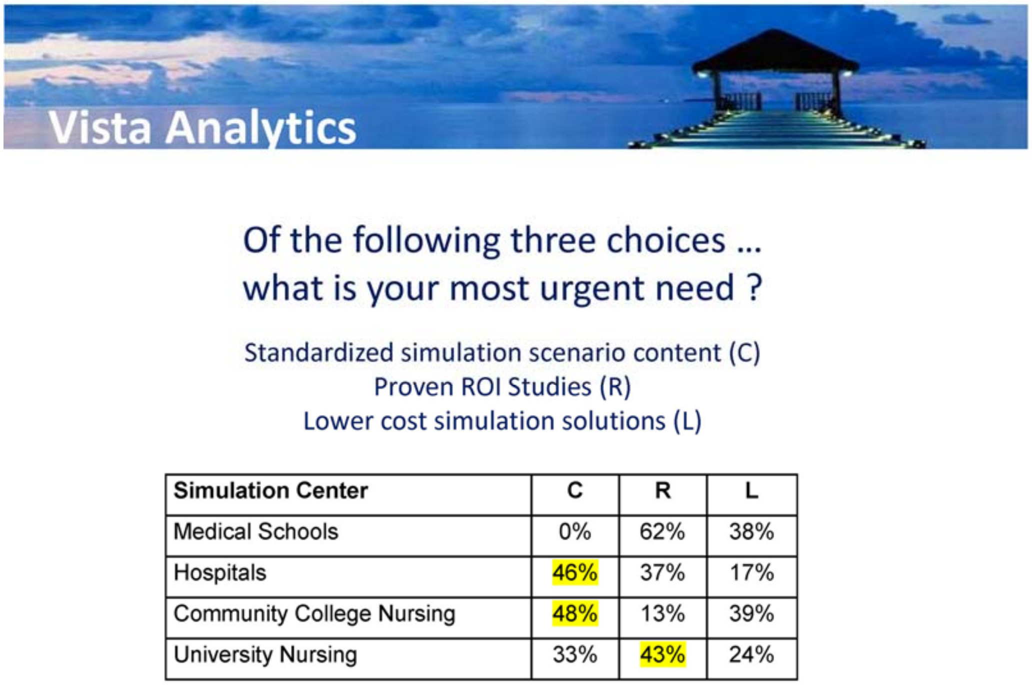


Vista Analytics

How are you managing your simulation center operations?

Paper based vs Using a collection of simulation management tools

Simulation Center	Paper based	Automated
Medical Schools	50%	50%
Hospitals	29%	71%
Community College Nursing	54%	46%
University Nursing	24%	76%

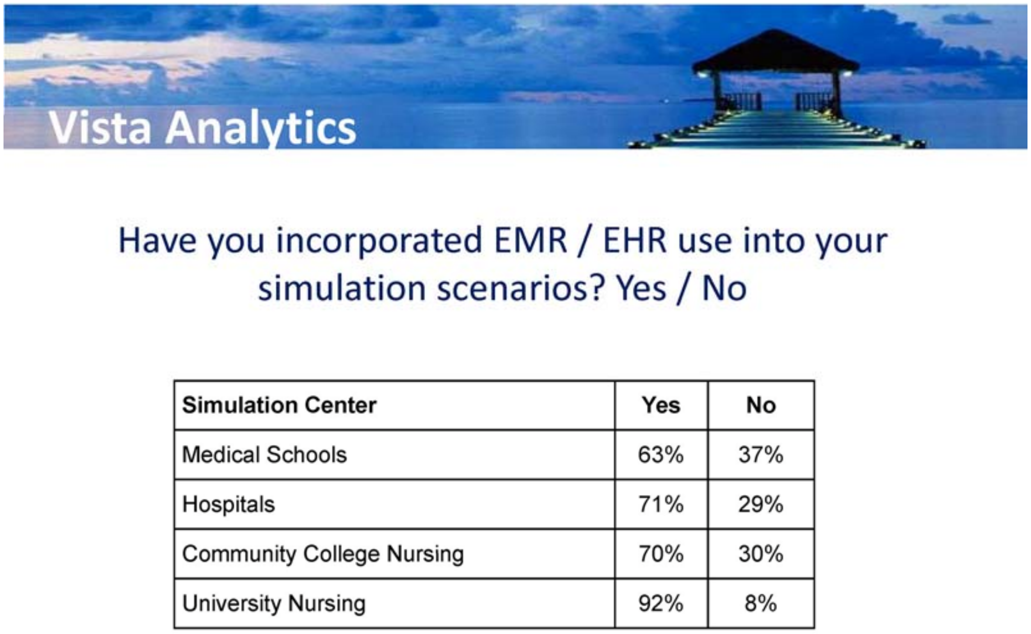


Vista Analytics

Of the following three choices … what is your most urgent need ?

Standardized simulation scenario content (C)
Proven ROI Studies (R)
Lower cost simulation solutions (L)

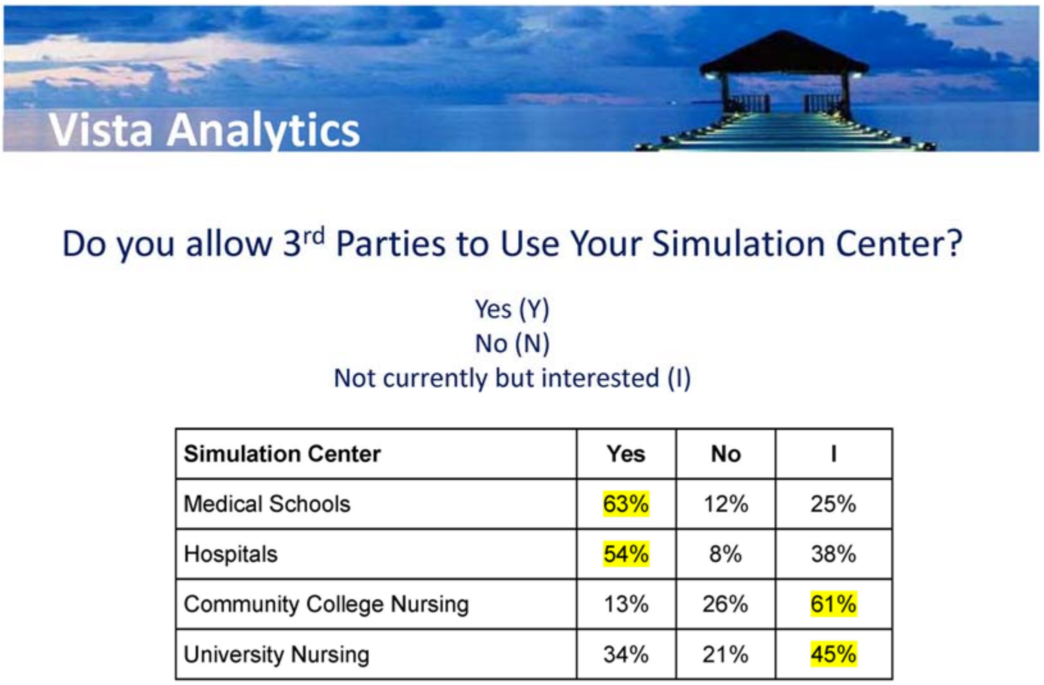
Simulation Center	C	R	L
Medical Schools	0%	62%	38%
Hospitals	46%	37%	17%
Community College Nursing	48%	13%	39%
University Nursing	33%	43%	24%



Vista Analytics

Have you incorporated EMR / EHR use into your simulation scenarios? Yes / No

Simulation Center	Yes	No
Medical Schools	63%	37%
Hospitals	71%	29%
Community College Nursing	70%	30%
University Nursing	92%	8%

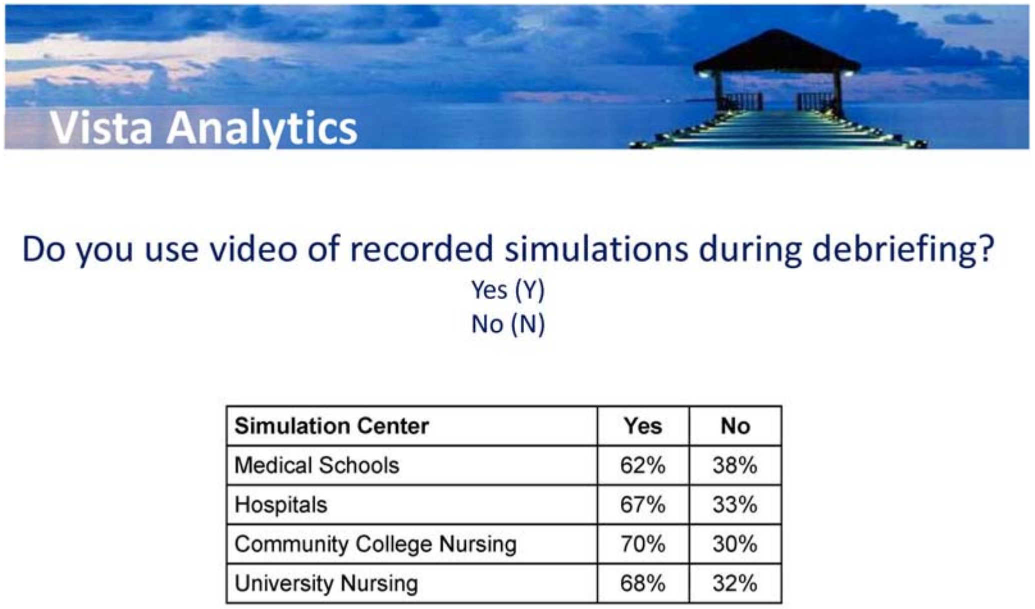


Vista Analytics

Do you allow 3rd Parties to Use Your Simulation Center?

Yes (Y)
No (N)
Not currently but interested (I)

Simulation Center	Yes	No	I
Medical Schools	63%	12%	25%
Hospitals	54%	8%	38%
Community College Nursing	13%	26%	61%
University Nursing	34%	21%	45%

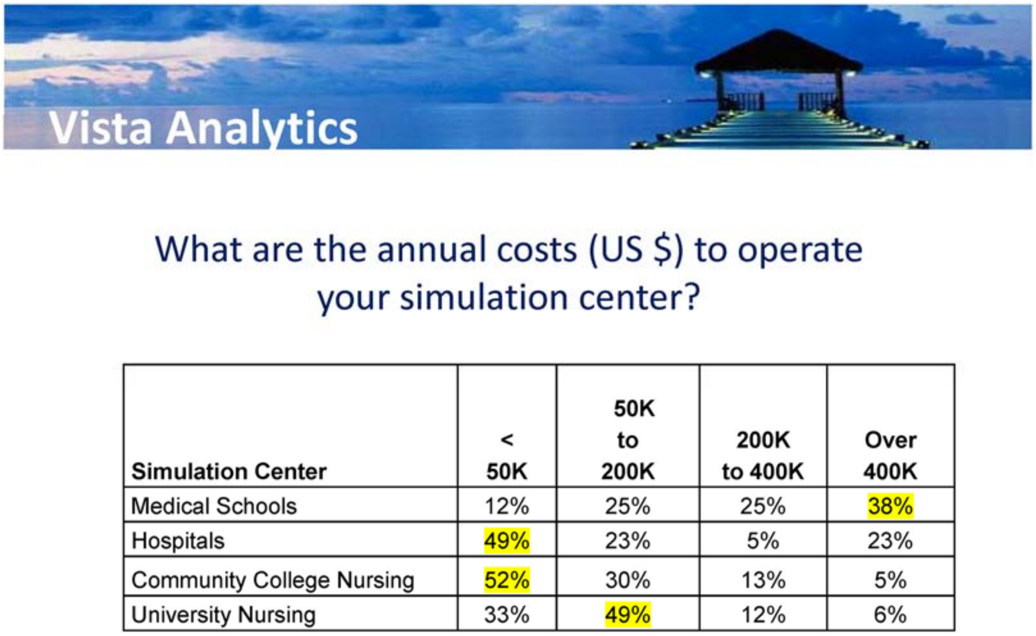


Vista Analytics

Do you use video of recorded simulations during debriefing?

Yes (Y)
No (N)

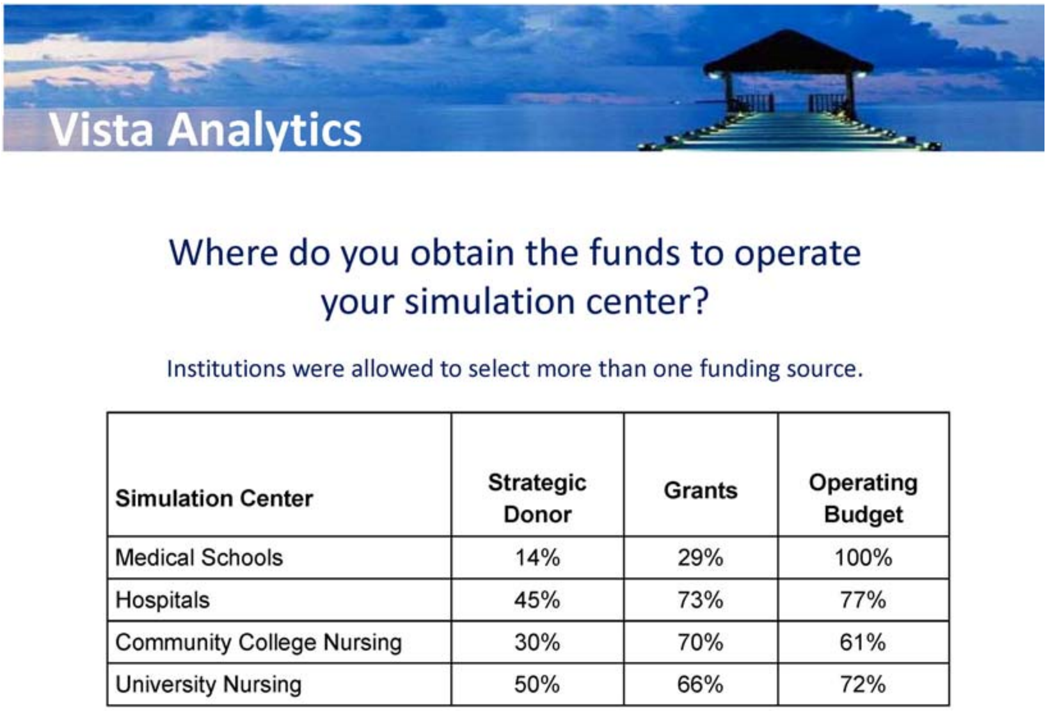
Simulation Center	Yes	No
Medical Schools	62%	38%
Hospitals	67%	33%
Community College Nursing	70%	30%
University Nursing	68%	32%



Vista Analytics

What are the annual costs (US $) to operate your simulation center?

Simulation Center	< 50K	50K to 200K	200K to 400K	Over 400K
Medical Schools	12%	25%	25%	38%
Hospitals	49%	23%	5%	23%
Community College Nursing	52%	30%	13%	5%
University Nursing	33%	49%	12%	6%



Vista Analytics

Where do you obtain the funds to operate your simulation center?

Institutions were allowed to select more than one funding source.

Simulation Center	Strategic Donor	Grants	Operating Budget
Medical Schools	14%	29%	100%
Hospitals	45%	73%	77%
Community College Nursing	30%	70%	61%
University Nursing	50%	66%	72%

Vista Analytics

Additional analysis results available at
www.vista-analytics.com/projects/medsimsurvey2010

Does perceived simulation center effectiveness vary by:
Simulation Center Size
Patients Models Used
Annual Budget
Years of Experience

The website referenced above will allow additional institutions to complete a survey representing their current operations.

Vista Analytics

Summary Comments

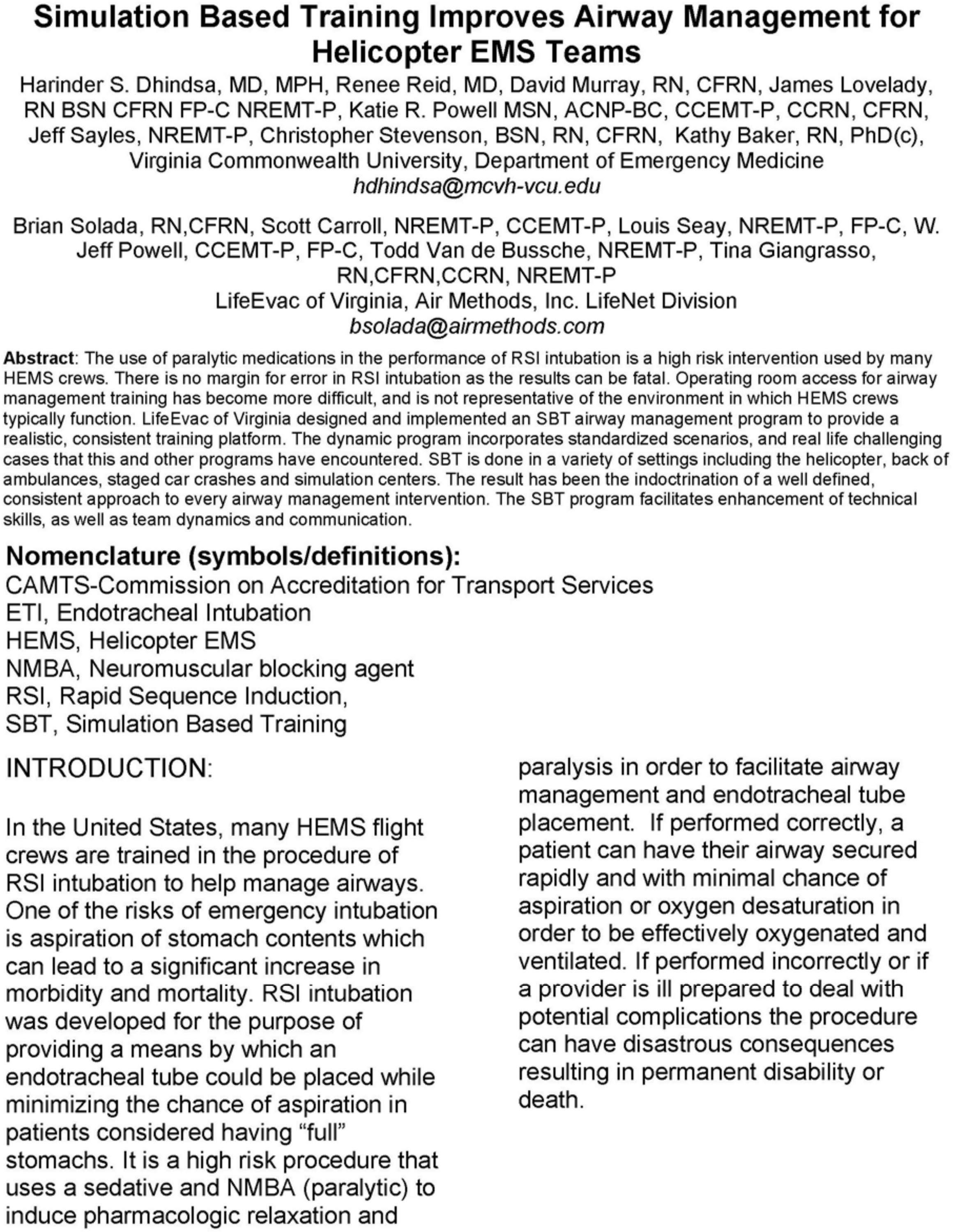
The survey collected what simulation centers are doing but not why they chose a specific approach. More insight into best practices would be useful.

Medical Schools are not adequately represented.
This group likely has a wealth of experience to share with other institutions.

Is perceived effectiveness closely related to simulation center quality?
Do we have an unbiased measurement of quality for a simulation center?
If we agree on these metrics we could start meaningful work on ROI studies.

It would be very useful to start collecting survey statistics on an annual basis to explore how the practice of medical simulation evolves over time.

3.2 Simulation Based Training Improves Airway Management for Helicopter EMS Teams



Simulation Based Training Improves Airway Management for Helicopter EMS Teams

Harinder S. Dhindsa, MD, MPH, Renee Reid, MD, David Murray, RN, CFRN, James Lovelady, RN BSN CFRN FP-C NREMT-P, Katie R. Powell MSN, ACNP-BC, CCEMT-P, CCRN, CFRN, Jeff Sayles, NREMT-P, Christopher Stevenson, BSN, RN, CFRN, Kathy Baker, RN, PhD(c),
Virginia Commonwealth University, Department of Emergency Medicine
hdhindsa@mcvh-vcu.edu

Brian Solada, RN,CFRN, Scott Carroll, NREMT-P, CCEMT-P, Louis Seay, NREMT-P, FP-C, W. Jeff Powell, CCEMT-P, FP-C, Todd Van de Bussche, NREMT-P, Tina Giangrasso, RN,CFRN,CCRN, NREMT-P
LifeEvac of Virginia, Air Methods, Inc. LifeNet Division
bsolada@airmethods.com

Abstract: The use of paralytic medications in the performance of RSI intubation is a high risk intervention used by many HEMS crews. There is no margin for error in RSI intubation as the results can be fatal. Operating room access for airway management training has become more difficult, and is not representative of the environment in which HEMS crews typically function. LifeEvac of Virginia designed and implemented an SBT airway management program to provide a realistic, consistent training platform. The dynamic program incorporates standardized scenarios, and real life challenging cases that this and other programs have encountered. SBT is done in a variety of settings including the helicopter, back of ambulances, staged car crashes and simulation centers. The result has been the indoctrination of a well defined, consistent approach to every airway management intervention. The SBT program facilitates enhancement of technical skills, as well as team dynamics and communication.

Nomenclature (symbols/definitions):

CAMTS-Commission on Accreditation for Transport Services
ETI, Endotracheal Intubation
HEMS, Helicopter EMS
NMBA, Neuromuscular blocking agent
RSI, Rapid Sequence Induction,
SBT, Simulation Based Training

INTRODUCTION:

In the United States, many HEMS flight crews are trained in the procedure of RSI intubation to help manage airways. One of the risks of emergency intubation is aspiration of stomach contents which can lead to a significant increase in morbidity and mortality. RSI intubation was developed for the purpose of providing a means by which an endotracheal tube could be placed while minimizing the chance of aspiration in patients considered having “full” stomachs. It is a high risk procedure that uses a sedative and NMBA (paralytic) to induce pharmacologic relaxation and paralysis in order to facilitate airway management and endotracheal tube placement. If performed correctly, a patient can have their airway secured rapidly and with minimal chance of aspiration or oxygen desaturation in order to be effectively oxygenated and ventilated. If performed incorrectly or if a provider is ill prepared to deal with potential complications the procedure can have disastrous consequences resulting in permanent disability or death.

BODY:

Historically the use of paralytic agents has been reserved for use by anesthesiologists and emergency physicians with years of specialized training in a hospital environment. However, the last decade has seen a proliferation of use of such agents in the out of hospital setting to facilitate airway management and endotracheal intubation by paramedics, nurses and respiratory therapists. The high acuity and severity of illness of patients typically transported by HEMS often places flight teams in the position of having to utilize advanced airway management skills such as RSI. HEMS teams are often called upon by emergency medical services providers for their expertise in airway management. Frequently they are required to utilize these skills on patients that require interfacility transport as well.

It is important to master the technical aspects of the RSI procedure as well as the associated risk assessment and decision making involved in the decision about whether or not to deploy the procedure. This level of risk assessment is essential in order to minimize any potential patient harm secondary to RSI intubation. SBT has become an increasingly popular tool for training in healthcare settings and has been specifically recommended for emergency medicine training [1-3].

LifeEvac of Virginia is a CAMTS accredited three base rotor wing program based in central Virginia that flies with a critical care paramedic and critical care RN. Although CAMTS accreditation is voluntary, it is rapidly evolving into an industry standard that symbolizes a commitment to quality and safety by transport agencies. CAMTS requires that flight crews participate in airway management training on patients that fall within the scope of care for the transport service (e.g., infants, children and adults) on a quarterly basis in order to ensure ongoing skill competency. The training standard states [4] "…no less than one successful live, cadaver, or mannequin intubation per quarter is required..." There is significant latitude left to the individual programs as to how they develop their educational programs to meet the standard. LifeEvac of Virginia developed and implemented a program utilizing SBT several years ago in order improve preparation for and success of RSI intubation and to minimize potential complications.

With any high risk procedure such as RSI it is essential that a consistent approach is utilized each time in order to maximize success and minimize complications. LifeEvac of Virginia developed a structured and guided SBT program for airway management built on the principals and algorithms endorsed in Emergency Airway Management [5]. Although there are many different ways to approach and perform RSI, this approach was chosen due to its reproducibility, ease of use, and due to the fact it is one of the most evidenced based approaches to airway management that currently exists.

The airway management training program is led by its Medical Director and a dedicated team of individually chosen crew members that have all taken The Difficult Airway Course: Emergency™, a physician level nationally offered emergency airway management course, directed by Ron Walls [6]. LifeEvac's airway instructor team is active year round in planning and designing training sessions. The airway instructors typically run three age based scenarios each quarter. Flight crew members rotate through each station in teams of two. (Figs 1, 2)

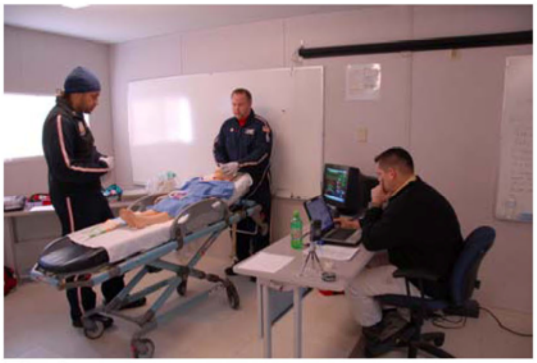


Fig 1

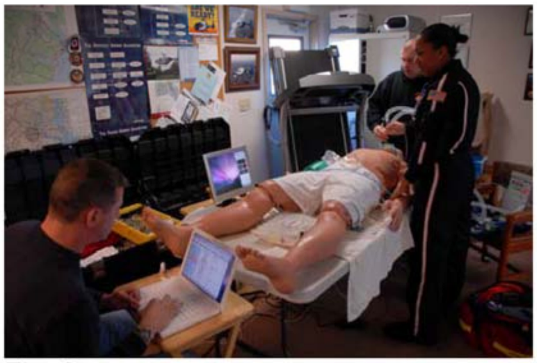


Fig. 2

They are provided an initial briefing of the scenario by the instructor and then given approximately thirty minutes to complete the scenario. The station is then debriefed with the crew and questions answered during the second thirty minutes. This same sequence is then repeated for the remaining two stations. These trainings take place over the course of two full days in order to cycle all of the crews through. At the end of the two days, the instructors debrief amongst themselves in order to identify any common trends or problems noted and to identify opportunities to improve the scenarios or overall training for the following quarter.

For each scenario, there are "critical actions" that reflect the fundamentals and essential technical aspects of RSI intubation as well as decision making and risk assessment for deciding whether or not to proceed with RSI intubation. All critical actions must be met in order to pass the scenario. If all criteria are not met, the team is remediated on the spot, and depending on the level of concern, may be brought back for additional remedial training at the discretion of the instructors and Medical Director. If a crew member's performance demonstrates numerous missed critical actions, their privilege to perform RSI intubation may be revoked until they demonstrate a satisfactory level of performance.

DISCUSSION:

One of the major advantages of SBT is that it has provided the ability to more realistically train HEMS crews in airway management in ways that were previously impractical. Traditionally many HEMS crews have done rotations in the operating room in order to gain intubation and airway management experience. Although these experiences are valuable, they are not reflective of the environment and conditions in which HEMS crews typically function. Operating rooms have good lighting, climate control, and patients are typically placed on the operating table in a position that is a comfortable level to intubate. The patients have empty stomachs (and thus are at low risk for aspiration), and there are plenty of resources and backup available should a complication arise. HEMS crews on the other hand are often called to manage airways in adverse conditions such as the middle of the night, in the rain, on the side of the road, or with patients trapped in vehicles, with poor lighting and intubation conditions, etc. All of their patients are considered to have "full stomachs" and are at high risk of aspiration. Should a complication arise, HEMS crews do not have the luxury of having backup resources available to assist. They have to be adequately prepared to manage the complications themselves. Another limitation of trying to send people to the operating room for training is that in recent years access to operating rooms has become more difficult due to concerns from anesthesiologists about liability exposure.

Various settings are utilized to train the crew in RSI intubation. These settings are intended to emulate environments that the flight crew may find themselves in, and include the back of an ambulance, a helicopter, the woods, vehicle entrapment, structure collapse, and in the hospital at the patient's bedside. (Fig 3)

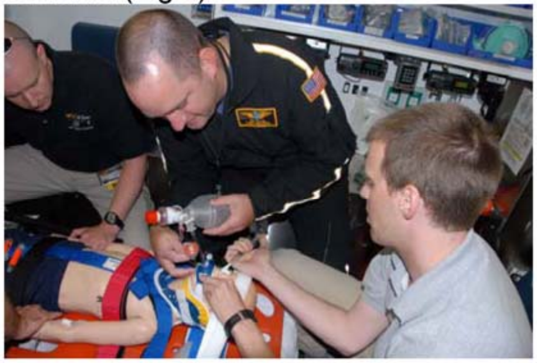


Fig 3

These various situations present different environmental, communication, and team work challenges. The flight crews have to manage the barriers to success these various environments may pose, while simultaneously implementing a consistent effective approach to airway management. There are no "critical actions" that are based on the environment in which airway management takes place. Rather these environments are selected in order to provide an additional level of realism and distraction in which crew members must demonstrate proficient airway management. (Figs 4-6)

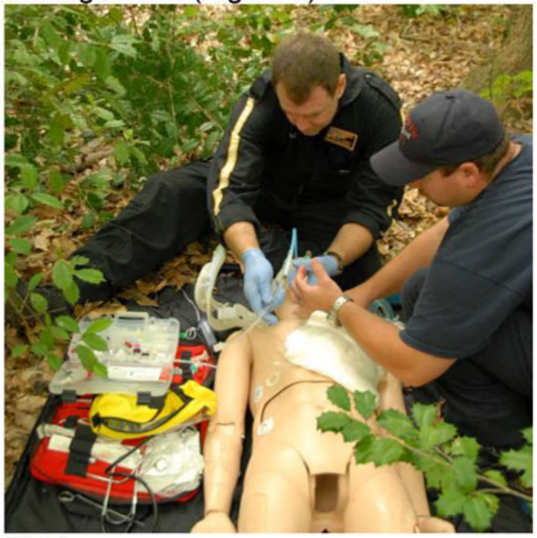


Fig 4

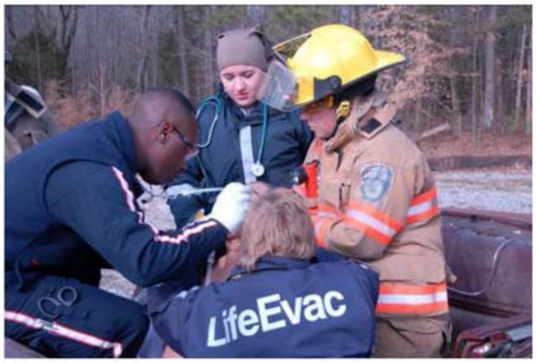


Fig 5

Fig 6

The scenarios are often based on actual calls that the program has experienced. Practicing in an environment with equipment that replicates reality raises situational awareness and increases the probability that the trained teamwork skills will transfer to real life practice. The advancement in technology that has resulted in portable human patient simulators has been instrumental in allowing us to provide training in austere environments. (Figs 7-8)

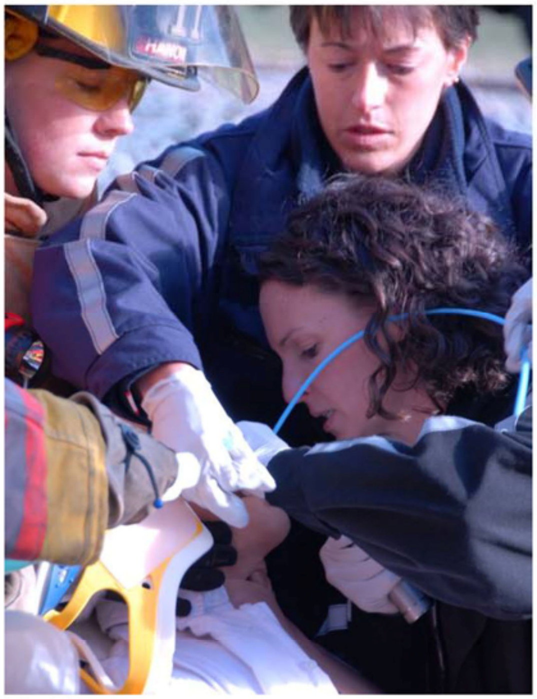


Fig 7

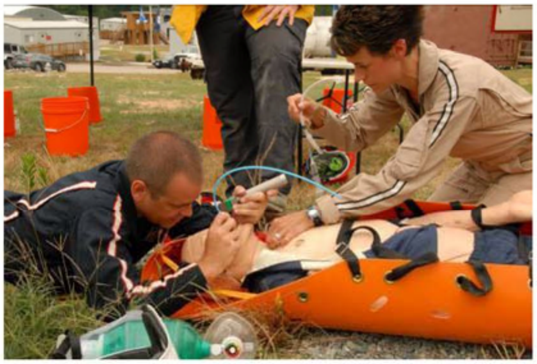


Fig 8

The dynamic human patient simulators offer several advantages over static manikins. (Fig.9)

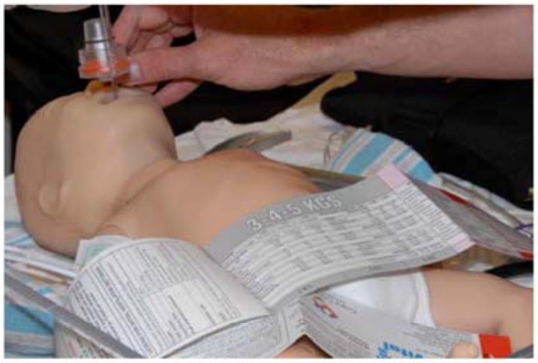


Fig 9

First, the instructors can set the difficulty of airway with regards to ability to visualize and access the vocal cords, by swelling the tongue or inducing jaw rigidity, and can create other physiologic complications. Physiologic responses to the flight crew's interventions as they work through the scenarios can be provided in real time. For example, if the crew does not adequately pre-oxygenate the patient, the instructor can cause the simulator to desaturate and become bradycardic in the middle of the procedure, forcing the crew to then have to manage this complication. Immediately after intubation, physiologic responses, such as lethal dysrhythmias that may occur can be reproduced with SBT and crew ability to respond to, diagnose and correct the problem measured.

CONCLUSION:

At LifeEvac of Virginia, the use of an SBT, using an active, dynamic human patient simulation training program for airway management and RSI intubation has led to high success rates of airway management with minimal complications. It has also improved medical team satisfaction with airway management training, in that we are able to more closely emulate real life situations.

The use of SBT as a training platform has allowed us to improve patient safety by being able to observe how team members anticipate, prepare for and manage potential complications associated with RSI intubation. Finally, use of the simulation based airway management program has reinforced the critical aspects of crew communication with respect to decision making and successful procedure implementation.

REFERENCES:

[1]Rosen, Michael A., et al. "Promoting teamwork: an event-based approach to simulation-based teamwork training for emergency medicine residents." Academic Emergency Medicine 15.11 (2008):1190-1198.

[2] Bond, William, et al. "The use of simulation in the development of individual cognitive expertise in emergency medicine." Academic Emergency Medicine 15.11 (2008):1037-1045.

[3] Small, S. D., et al. "Demonstration of high-fidelity simulation team training for emergency medicine." Academic Emergency Medicine 6.4 (1999):312-323.

[4]Commission on Accreditation for Transport Services (CAMTS) Accreditation Standards, 7th Edition Accreditation Standards. http://www.camts.org/component/option,com_docman/task,cat_view/gid,17/Itemid,44/

[5] Manual of Emergency Airway Management. Edited by Ron M. Walls, Michael F. Murphy, and Robert C. Luten 3rd ed, Philadelphia, PA, Wolters Kluwer/Lippincott Williams & Wilkins, 2008 .ISBN-13: 978-0-7818-8494-8

[6] The Difficult Airway Course: Emergency™. The Airway Site: The Definitive Site for Airway Management. http://www.theairwaysite.com/pages/page_content/Airway_Emergency_More.aspx

3.3 Leveraging Game Consoles for the Delivery of TBI Rehabilitation

Leveraging Game Consoles for the Delivery of TBI Rehabilitation

Taryn Cuper, Thomas Mastaglio
MYMIC LLC
taryn.cuper@mymic.net, thomas.mastaglio@mymic.net

Yuzhong Shen
Old Dominion University
yshen@odu.edu

Robert Walker, MD
Eastern Virginia Medical School
walker@evms.edu

Abstract. Military personnel are at a greater risk for traumatic brain injury (TBI) than the civilian population. In addition, the increase in exposure to explosives, i.e., improvised explosive devices, in the Afghanistan and Iraq wars, along with more effective body armor, has resulted in far more surviving casualties suffering from TBI than in previous wars. This effort presents the results of a feasibility study and early prototype of a brain injury rehabilitation delivery system (BIRDS). BIRDS is designed to provide medical personnel treating TBI with a capability to prescribe game activities for patients to execute using a commercially available game console, either in a clinical setting or in their homes. These therapeutic activities will contribute to recovery or remediation of the patients' cognitive dysfunctions. Solutions such as this that provide new applications for existing platforms have significant potential to address the growing incidence of TBI today.

1.0 INTRODUCTION

Military personnel are at a greater risk for traumatic brain injury (TBI) than the civilian population. In addition, the increase in exposure to explosives, i.e., improvised explosive devices, in the Afghanistan and Iraq wars, along with more effective body armor, has resulted in far more surviving casualties suffering from TBI than in previous wars. For example, 14-18% of Vietnam veterans had a brain injury while from 2003-2005, Walter Reed Army Medical Center reported 31% of those combat casualties admitted had a brain injury [1]. Such injuries can leave Service members physically scarred and cognitively challenged.

2.0 BACKGROUND

The research team, comprised of members from MYMIC LLC, Old Dominion University, and Eastern Virginia Medical School, designed an intervention for individuals with TBI as a result of a Phase I SBIR award. The Brain Injury Rehabilitation Delivery System (BIRDS) is designed to deliver rehabilitation games via a commercial game console and assess performance over time from a baseline measurement throughout treatment. The vision for BIRDS is to provide medical personnel treating TBI a capability to prescribe game activities for patients to execute using that game console, either in a clinical setting or in their homes. These therapeutic activities will contribute to recovery or remediation of the patients' cognitive dysfunctions. BIRDS design is initially focused on TBI resulting from combat; however, our vision includes its application for treating injuries resulting from other activities (e.g., sports or accidents) and age-related dysfunctions. This paper details our research findings and described the implementation of the BIRDS concept prototype.

3.0 DESIGN

3.1 The Microsoft Xbox 360 platform

Microsoft Xbox 360 is the second-generation game console released by Microsoft. It supports multiple interface devices, such as the keyboard, joystick, and racing wheel Both wired and wireless controllers, Figures 1(a) and (b), provide user interaction via buttons, thumb sticks, triggers, and a directional pad. The joystick, Figure 1(c) also provides these options, though in a different configuration and with greater emphasis on the joystick function. The racing wheel devices, Figures 1(d) and (e), can track both hand, arm, and foot motions as well as hand-foot coordination.

The freely available Microsoft XNA Game Studio is an integrated environment for game development for Xbox 360, personal computer, and the handheld Microsoft Zune Player XNA Game Studio consists of two major components: XNA Framework and a set of tools and templates to facilitate rapid game development. The XNA Framework is a set of optimized cross-platform libraries for game development based on the Microsoft .NET Framework Technology. It encapsulates low-level details involved in developing games and allows game developers to focus more on the content and high-level gaming experience.

3.2 Mapping platform capabilities to TBI rehabilitation skills foci

We mapped the capabilities of the Xbox 360 to a selection of cognitive and motor skills foci in clinical TBI rehabilitation settings. Based on these data points, we selected a limited set of skills to be addressed in the concept prototype.

The features and capabilities of both Xbox360 hardware and software that are associated with the interface between the system and the human were analyzed with respect to their applicability to TBI rehabilitation foci, including cognitive and fine or gross motor functions. Table 1 contains these mappings, as well as some example task suggestions on how to realize skill rehabilitation within multiple modalities.

In this research we identified the joystick as a promising interface for the intended user population because of its ability to support gross motor, as well as fine motor movements and cognitive skills. For example, a patient lacking fine motor capability (e.g. controlled finger movements) can engage larger muscle groups, i.e. arm muscles, to work the joystick. However, while Xbox360 itself supports different input devices, including the joystick, Microsoft XNA Game Studio currently supports only game controllers, keyboard, and mouse. Therefore, the BIRDS concept prototype was developed for the game controller device. This device can support both fine motor and cognitive skills. Future goals will include adding support for using gross motor skills in BIRDS.

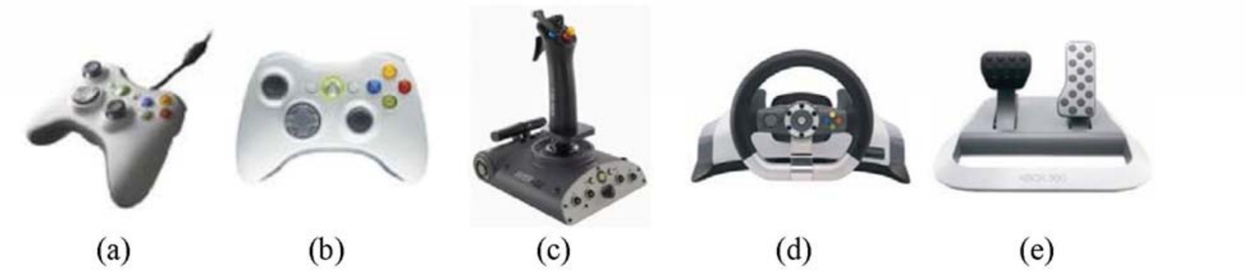


Figure 1. Xbox 360 input devices. (a) Wired controller. (b) Wireless controller. (c) Joystick. (d) Racing wheel. (e) Pedal.

Table 1. Mapping of Xbox interface capabilities to applicable TBI rehabilitation skills foci.

Xbox capabilities		Options	Rehab skills foci	Potential tasks
Interfaces	Wired game controller	Buttons - press Joystick - push / pull	Fine motor	- Move object to desired/directed location (large to small objects) - "Thread" a needle through different sized holes (large to small holes) - Coordinate movement (joystick) with selection (button)
	Wireless game controller			
	Joystick	Button - press Joystick - push / pull	Gross Motor Fine Motor	
	Racing wheel	Wheel - turn Buttons - press	Gross Motor	
	Racing pedals	Pedals - press (feet)	Gross Motor	
	Software	Gameplay	Cognitive	
			- Attention	- Introduce stimuli during task (i.e. problem solving)
			- Memory	- Recall objects from the previous screen
			- Information Processing	- Determine which 3 of 4 objects are similar
			- Apraxia	- Sequential instructions

3.3 Identifying performance metrics

We further identified performance metrics needed for clinicians to assess each TBI rehabilitation skill in a meaningful way. This step is especially important as an objective measure of progress and as such, should be supported by existing evidence-based scales, and be relevant to patient functional outcomes. In addition, performance metrics will play an important role in the validation of any computer/console based cognitive and motor skills rehabilitation tool.

In order to do this, standard clinical protocol for evaluating both motor and cognitive skills in patients was compared with the capabilities of the Xbox360 to determine the applicable performance metrics relative to TBI rehabilitation skills.

The BIRDS concept supports objective performance assessment, as opposed to a more subjective method of determining patient progress, for example, an evaluation such as, “The patient *appears* to be responding more quickly to commands.” With the aid of an SME, we determined performance metrics for the TBI rehabilitation skills that were previously identified. Table 2 details these metrics.

Table 2. Performance metrics for applicable TBI rehabilitation skills foci.

Xbox360 Interface		Rehab skills foci	Performance metric
Wired game controller	Buttons - press Joystick - push / pull	Fine motor	Accuracy Accuracy (spatial) Completion time
Wireless game controller			
Joystick	Button - press Joystick - push / pull	Gross Motor Fine Motor	Accuracy (spatial) Completion time
Racing wheel	Wheel - turn Buttons - press	Gross Motor	Accuracy (spatial) Completion time
Racing pedals	Pedals - press (feet)	Gross Motor	Accuracy Completion time
Software	Gameplay	Cognitive	
		- Attention	Accuracy Reaction Time
		- Memory	Accuracy
		- Information Processing	Accuracy Reaction Time
		- Apraxia	Accuracy Reaction Time Sequence

4.0 DEVELOPMENT OF THE PROTOTYPE

The BIRDS concept prototype (Figure 2) was developed with a focus on cognitive skills and a secondary requirement of fine motor movement. The following cognitive skills were selected with the goal of developing a concept prototype for a minimum of two of these skills. These particular skills were selected because of their foundational role in cognition and information processing. Each of these low-level skills are required if a person is to function normally and/or gain functional independence following brain injury.

- Attention (cognitive): The ability to focus selectively on a specific stimulus, to maintain that focus, and to shift that focus at will.
- Memory (cognitive): The ability to store, retain, and recall information.
- Apraxia (cognitive with some additional fine motor requirements): the ability to follow a sequence of instructions

Each game has two levels of difficulty: easy and hard. In the Attention game, the player will observe one letter of the English alphabet at a time and will need to respond when the letter 'A' appears by pressing the 'A' button of the Xbox 360 controller. At the easy level, letters are displayed in a consistent location, one after another. At the hard level, distracting graphics are included and displayed, such as stars appearing and disappearing in different sizes and in different locations, while the letters themselves are displayed in random locations. In addition, easy and hard are further differentiated by the length of time each stimulus is displayed. The purpose of the game is to test and improve the player's ability to concentrate on a specific, randomly displayed stimulus.

In the Memory game, the player will first see an image, see Figure 3(a), with the goal of memorizing the image. Then the player will see four images, see Figure 3(b), including the original image, and must select which one matches the original image. Easy and hard levels are determined by the amount of time the original image is displayed. Images from the following five categories are used: animals, food, plants, transport, and weapons.

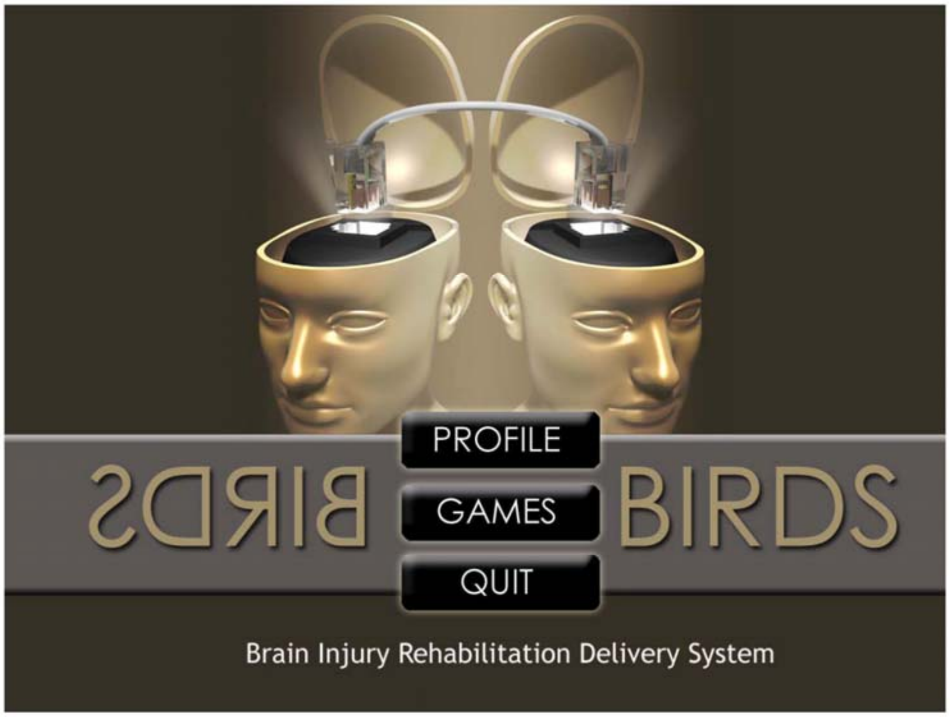


Figure 2. BIRDS concept prototype screen mock-up.

Each group contains more than 10 images. During each play, an image from a certain group is displayed first; then three other images from the same group are displayed in addition to the original image. In Phase II, testing will determine if the player has varied performance for different types of images, e.g., Food or People.

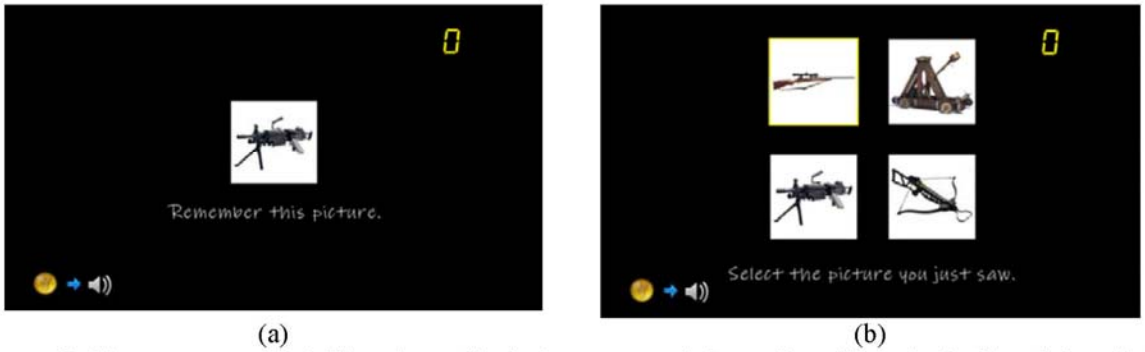
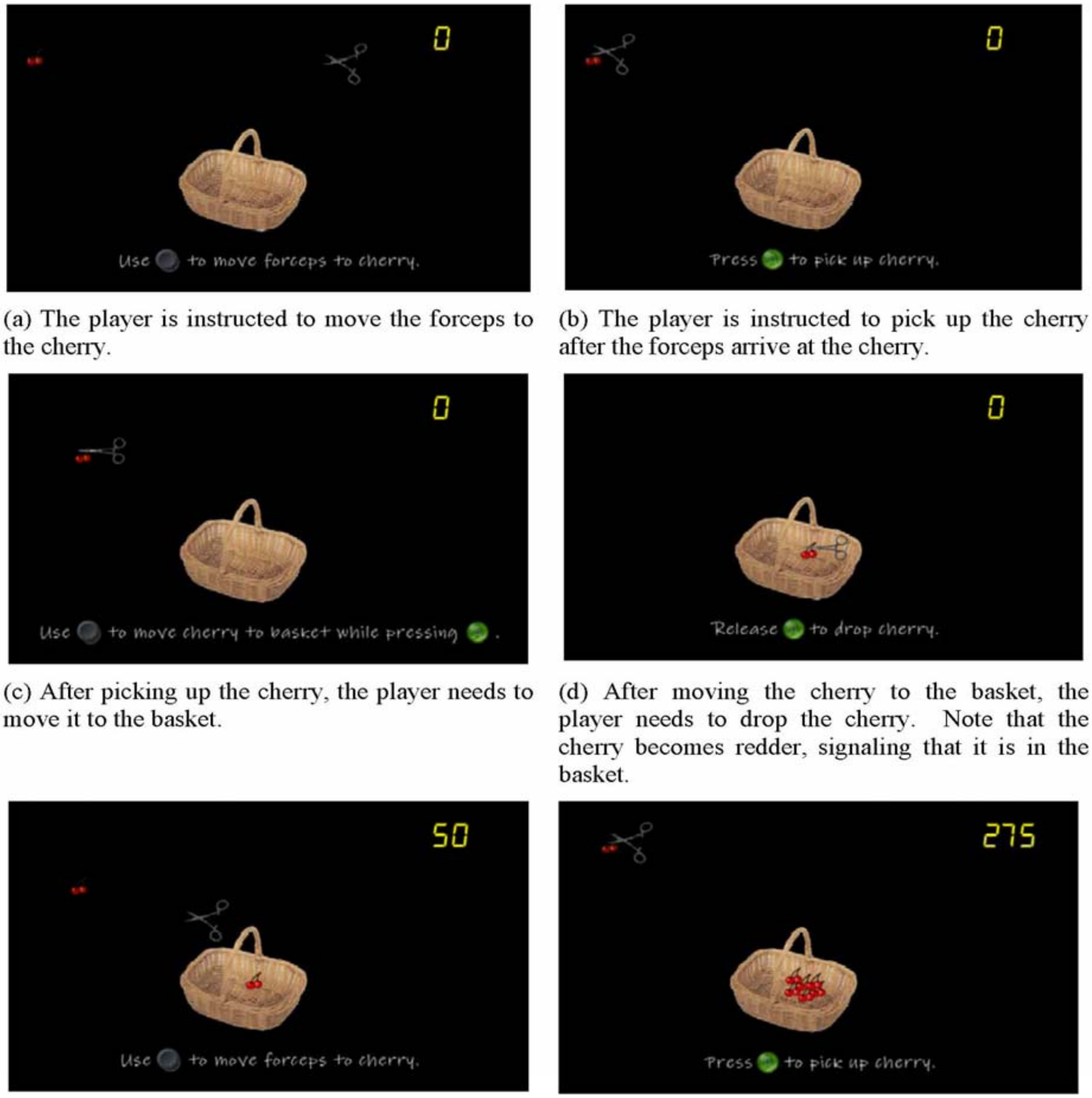


Figure 3. Memory game. (a) The player first observes a picture; then (b) selects the picture he/she just saw.

The apraxia game will require the player to follow on screen instructions (and/or instructions given by a virtual trainer) to complete a sequence of actions. The player will need to use both hands to manipulate several controls (e.g., button, thumb stick) at the same time in order to operate an on-screen instrument to execute the required task(s). For example, as shown in Figure 4, the player needs to use the left thumb stick to move the forceps to the cherry, Figure 4(a). After the forceps arrive at the cherry, the player presses button A using right hand, Figure 4(b). Then the players uses the thumb stick (left hand) to move the cherry to the basket, while pressing button

A (right hand), Figure 4(c). Finally, the player releases button A to drop the cherry into the basket, Figure 4(d).



Use ● to move forceps to cherry.

(a) The player is instructed to move the forceps to the cherry.

Press ● to pick up cherry.

(b) The player is instructed to pick up the cherry after the forceps arrive at the cherry.

Use ● to move cherry to basket while pressing ●.

(c) After picking up the cherry, the player needs to move it to the basket.

Release ● to drop cherry.

(d) After moving the cherry to the basket, the player needs to drop the cherry. Note that the cherry becomes redder, signaling that it is in the basket.

Use ● to move forceps to cherry.

(e) After the cherry is dropped in the basket, another cherry appears to begin a new play.

Press ● to pick up cherry.

(f) A number of cherries have been collected.

Figure 4. Apraxia game (cherry picking).

The difficulty level of the apraxia (cherry picking) game is determined by the time allowed to complete the steps needed to pick up the cherry; in addition, the location of the basket is not fixed and the starting positions of the forcepts and cherries are also randomized to require different motions and avoid repetitous gameplay.

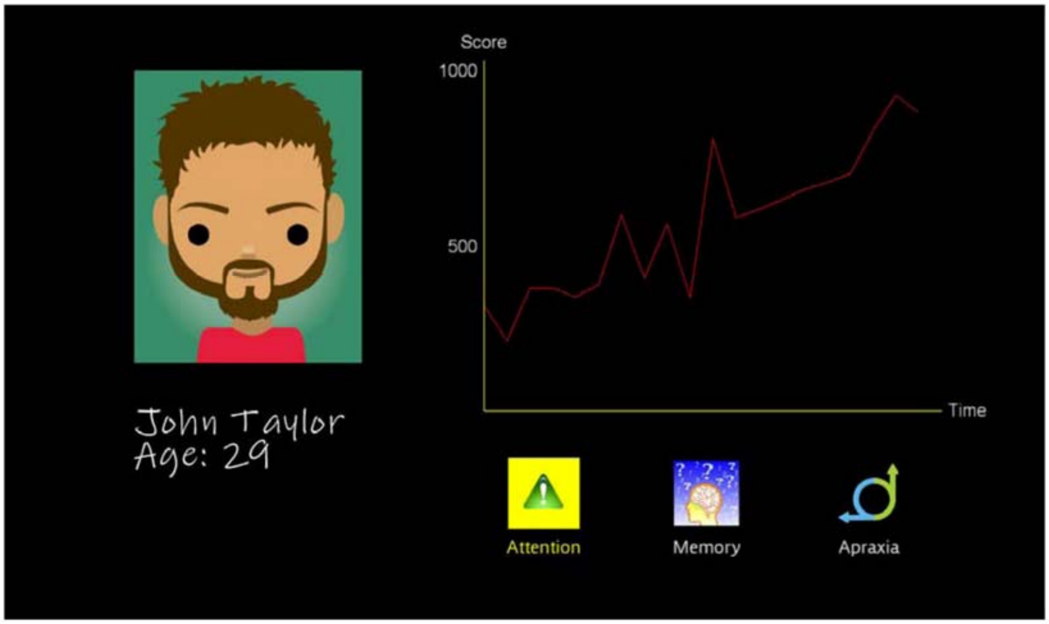


Figure 5. Example patient profile screen.

Users can view their overall progress in a given intervention by viewing their profile. Figure 7 shows a sample patient profile screen with actual progress graphed over time. Viewing their profile gives the patient the opportunity to see objectively how they are progressing, which mirrors the feedback given by physiatrists who periodically assess their patients' capabilities. It also provides a snapshot view of progress to the physician.

5.0 CONCLUDING REMARKS

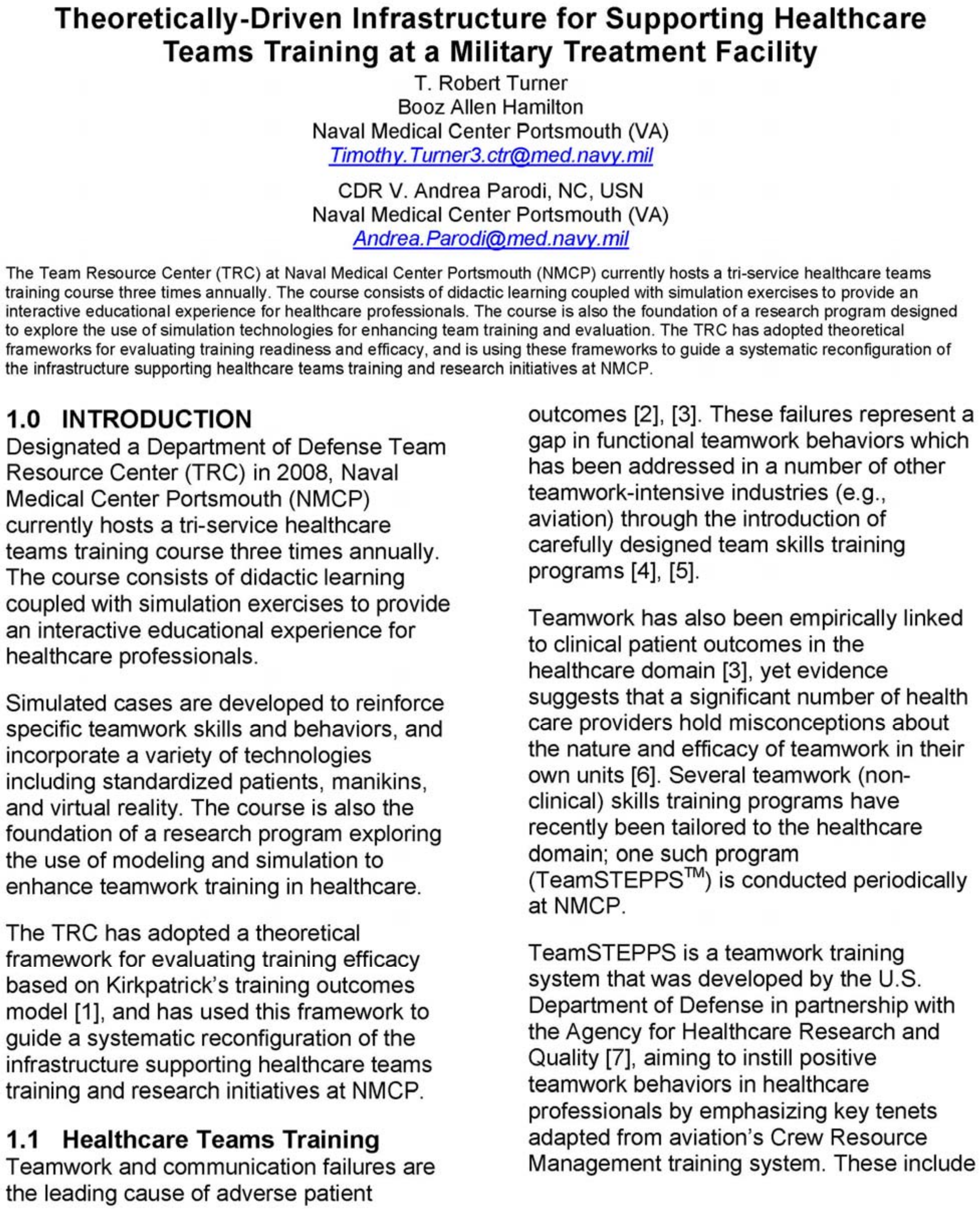
The key research findings described in this paper provide a foundation for the development of a videogame console-based system for TBI-related cognitive and motor skills rehabilitation. Such a tool may provide the foundational capabilities that are required for a patient to achieve functional independence in his or her everyday life. In addition, it has the potential to evolve to support higher order cognitive skills, e.g. information processing, and broader motor skills as well. Importantly, patient therapy can then be standardized and progress tracked objectively, providing a rich breeding ground for research in TBI rehabilitation, enriching the field and potentially resulting in even more effective treatments. Looking ahead, in developing a cognitive and motor skills rehabilitation console-based tool, attention must also be focused on the advances in available technology. For example, product releases such as Project Natal hold significant potential for the future extension and applicability of BIRDS treatment. Such technology could address physical as well as cognitive rehabilitation, providing benefits to a wider range of patients and more closely mimicking a physician's evaluation of a patient.

This work was supported by the US Army Medical Research and Materiel Command under Contract No. W81XWH-09-C-0127. The views, opinions, and/or findings contained in this report are those of the authors and should not ne construed as official Department of the Army position, policy or decision unless so designated by other documentation.

References

[1] Adams, GJ. Incidence of Traumatic Brain Injury in The Military. *EzineArticles.com*. 2007. Available at: http://ezinearticles.com/?Incidence-Of-Traumatic-Brain-Injury-In-The-Military&id=752358.

3.4 Theoretically-Driven Infrastructure for Supporting Healthcare Teams Training at a Military Treatment Facility



Theoretically-Driven Infrastructure for Supporting Healthcare Teams Training at a Military Treatment Facility

T. Robert Turner
Booz Allen Hamilton
Naval Medical Center Portsmouth (VA)
Timothy.Turner3.ctr@med.navy.mil

CDR V. Andrea Parodi, NC, USN
Naval Medical Center Portsmouth (VA)
Andrea.Parodi@med.navy.mil

The Team Resource Center (TRC) at Naval Medical Center Portsmouth (NMCP) currently hosts a tri-service healthcare teams training course three times annually. The course consists of didactic learning coupled with simulation exercises to provide an interactive educational experience for healthcare professionals. The course is also the foundation of a research program designed to explore the use of simulation technologies for enhancing team training and evaluation. The TRC has adopted theoretical frameworks for evaluating training readiness and efficacy, and is using these frameworks to guide a systematic reconfiguration of the infrastructure supporting healthcare teams training and research initiatives at NMCP.

1.0 INTRODUCTION

Designated a Department of Defense Team Resource Center (TRC) in 2008, Naval Medical Center Portsmouth (NMCP) currently hosts a tri-service healthcare teams training course three times annually. The course consists of didactic learning coupled with simulation exercises to provide an interactive educational experience for healthcare professionals.

Simulated cases are developed to reinforce specific teamwork skills and behaviors, and incorporate a variety of technologies including standardized patients, manikins, and virtual reality. The course is also the foundation of a research program exploring the use of modeling and simulation to enhance teamwork training in healthcare.

The TRC has adopted a theoretical framework for evaluating training efficacy based on Kirkpatrick's training outcomes model [1], and has used this framework to guide a systematic reconfiguration of the infrastructure supporting healthcare teams training and research initiatives at NMCP.

1.1 Healthcare Teams Training

Teamwork and communication failures are the leading cause of adverse patient outcomes [2], [3]. These failures represent a gap in functional teamwork behaviors which has been addressed in a number of other teamwork-intensive industries (e.g., aviation) through the introduction of carefully designed team skills training programs [4], [5].

Teamwork has also been empirically linked to clinical patient outcomes in the healthcare domain [3], yet evidence suggests that a significant number of health care providers hold misconceptions about the nature and efficacy of teamwork in their own units [6]. Several teamwork (non-clinical) skills training programs have recently been tailored to the healthcare domain; one such program (TeamSTEPPS^{TM}) is conducted periodically at NMCP.

TeamSTEPPS is a teamwork training system that was developed by the U.S. Department of Defense in partnership with the Agency for Healthcare Research and Quality [7], aiming to instill positive teamwork behaviors in healthcare professionals by emphasizing key tenets adapted from aviation's Crew Resource Management training system. These include communication, team structure, leadership, mutual support, and situation monitoring.

The NMCP TeamSTEPPS program is designed to provide students with the skills and tools necessary for effective teamwork, as well as hands-on skills training via simulation. The key assumptions are that critical teamwork skills are reinforced during the training program and that patient outcomes will improve as a result of these skills being transferred to the work environment.

However, recent research on healthcare team training efficacy has yielded mixed results [8], [4], [5]. One possible reason for this is the complexity of linking team performance characteristics to measurable outcomes. Few healthcare team training initiatives currently implement a comprehensive evaluation protocol, thus failing to demonstrate the achievement of intermediate training objectives. We have developed a multi-level assessment protocol for healthcare team training outcomes, which includes behavioral observation and analyses stemming from training scenarios conducted at the NMCP Healthcare Simulation Center.

1.2 Simulation as a Tool for Healthcare Teams Training

Simulation and teamwork are two relatively novel aspects of healthcare training that have only recently begun to receive significant attention. Using a simulated operating theater to examine surgical skill acquisition and maintenance over time, Moorthy et al. [9] discovered that communication skills (unlike technical skills) do not develop naturally as a result of increased job experience. Rather, these skills must be consciously trained and reinforced. Further, the ability of healthcare professionals to accurately and reliably assess their own non-technical performance is not sufficient to promote self-regulation and skill acquisition [10]. Effective teamwork is critical for patient safety, yet becoming an expert team member requires practice. Evidence is beginning to emerge in support of simulation as an ideal tool for healthcare teams training [11], [4], [12]. At NMCP, we have been able to successfully integrate simulation into our healthcare teams training program with the goal of enhancing teamwork skill acquisition through hands-on practice.

1.3 Early Training Infrastructure and Resource Availability at NMCP

1.3.1 <u>Audio/Video Data Support</u>

The NMCP Healthcare Simulation Center is equipped with a network of ceiling-mounted video cameras and microphones which feed into a central control area for programming camera angles, recoding and storing audio/video data, and rendering hard-copy discs. This network covers the entire center, with several cameras and microphones dedicated to each training room throughout the center. For training purposes, it is often desirable to play back the video feed so that participants have a chance to review performance and debrief individually or with instructors.

1.3.2 <u>Patient Simulators</u>

The Simulation Center houses a number of manikin simulators, including the Laerdal SimMan 3G, Laerdal SimBaby NEWBIE, and METI iStan. Training scenarios for the manikins can be created and delivered using either PC or Mac systems, depending on user needs. A larger number of part-task skills and box trainers are also available, covering a range of tasks such as IV line insertion, orthroscopy, central line placement, airway management, suturing, lumbar puncture, and more.

1.3.3 <u>Virtual Reality</u>

To support the acquisition and maintenance of psychomotor skills for minimally invasive surgery (MIS), the Simulation Center offers a variety of virtual reality (VR) trainers. These VR trainers can be used practice upper and lower GI respiratory scope insertion, colonoscopy, hernia repair, laparoscopic cholecystectomy, and more. Each of the VR systems provides students with a set of physical instruments that must be manipulated in order to affect some physiological change on a virtual patient. These changes are fed back to the student through display of realistic sights and sounds (delivered through digital monitors) and touch feedback (delivered through the instruments themselves).

1.3.4 Standardized Patients

The Team Resource Center at NMCP currently partners with the Theresa A. Thomas Professional Skills Teaching and Assessment Center at the Eastern Virginia Medical School to provide standardized patients in support of our training courses. Standardized patients are highly-trained actors who present symptoms of illness and then assess healthcare professionals' diagnostic and interpersonal skills during face-to-face interactions. The standardized patients are carefully integrated into training scenarios to produce a realistic social context for students to practice within.

1.4 Performance Assessment: Kirkpatrick Training Model

The NMCP Healthcare Simulation Center offers a range of simulation technologies to support team training. However, technology alone is not the key to training success. Simulation must be part of a larger training process, including a well-designed curriculum and evaluation protocol. The latter is perhaps more often neglected than the former [4], [5]. TeamSTEPPS students at NMCP are evaluated throughout the course along four levels of measurable outcome: reaction, learning, behavior, and results. These levels are based on Kirkpatrick's [1] model of training outcomes assessment.

1.4.1 Level 1: Reaction

Reaction-level feedback reflects the degree to which the training course and its content are valued by the students. This generally consists of asking students to complete a short pencil-and-paper feedback survey at the end of the course. Reaction-level data also help course administrators identify program strengths and opportunities for improvement.

1.4.2 Level 2: Learning

Learning-level feedback represents the degree to which relevant student attitudes and knowledge are positively impacted by participation in the course. Instruments such as the Team Attitudes Questionnaire [13] and modified TeamSTEPPS knowledge assessment instruments [7] are administered before training to establish baseline metrics. Upon completion of the course, students are asked to complete these assessments a second time to establish a comparison metric. The instruments may be continually administered over time to determine whether teamwork attitudes and knowledge have been sustained.

1.4.3 Level 3: Behavior

Behavioral outcomes reflect the degree to which core TeamSTEPPS tools and strategies have been successfully implemented in hospital units. To generate this form of data, trained observers spend time in the units monitoring and recording team activity using behavioral checklists. Behavior-level feedback is also generated during the training course, when students are asked to apply TeamSTEPPS concepts to resolve simulated case scenarios. These simulation sessions are audio/video recorded, and performance is critiqued during a post-scenario debrief.

1.4.4 Level 4: Results

Unit-specific metrics are maintained on a unit-by-unit basis and are analyzed periodically by the TRC. These metrics may reflect patient outcome data, procedural checklists, brief/debrief content analysis, and a number of other teamwork-related evaluations. Results-level outcomes reflect the organizational impact of the TeamSTEPPS training program over time.

1.5 TRC Performance and Assessment Needs

The first two outcome levels are assessed with pencil-and-paper survey instruments designed to record students' perceptions, knowledge and attitudes toward team training. Behavioral outcomes (Level 3) involve demonstration of acquired skill through hands-on TeamSTEPPS implementation. This is unlikely to occur in the work setting unless students are provided sufficient practice and feedback during training.

Carefully designed simulation scenarios allow students to practice using TeamSTEPPS skills and strategies in a safe learning environment and to receive feedback from colleagues and instructors so that these skills can be reinforced. However, conducting team training scenarios and video debrief sessions for TeamSTEPPS was not originally possible due to training infrastructure incompatibility.

Early in the training program it was determined that the Simulation Center's audio/video network was not designed to support teamwork debriefing. Rather, the ceiling-mounted video cameras and audio devices were installed to provide top-down, patient-centered perspectives for evaluating clinical proficiency. The cameras and microphones themselves produced low-grade surveillance quality sound and imaging. Further, the computer system dedicated to rendering hard-copy discs of the audio/video data for the purposes of analysis and debriefing required several hours to process, making immediate training debriefs impractical.

In addition to training debriefs, high-quality audio/video data were necessary to train unit and ward observers, to analyze effectiveness of training scenarios, and to demonstrate TeamSTEPPS skill improvement over a number of trials. Aside from the training center's infrastructure incompatibility for TeamSTEPPS, it was also determined that our conventional teamwork simulation scenarios were not producing the desired learning effects.

Our initial approach to scenario development was to embed specific TeamSTEPPS learning objectives into a series of patient-centered clinical scenarios, with roles and learning opportunities available for all members of the healthcare team. The goal was to provide for clinical fidelity at the highest possible level, thus allowing students to focus on improving teamwork rather than becoming distracted by an unfamiliar technical context (e.g., lack of functioning anesthesia machine, absence of an attending physician, or varying the point at which an official timeout is conducted before surgery). However, it quickly became evident that our strong emphasis on clinical detail was counterintuitive to our goals of delivering quality non-clinical training scenarios.

The majority of TeamSTEPPS students were from outside NMCP and were registered to attend by TRICARE Management Activity (TMA). Frequently we did not know the students' background information (i.e., job role, specialty, training needs, etc.) in advance. As a result, simulation scenarios targeting students with specific job roles and skill requirements proved to be too inflexible and were difficult to manage from an administrative perspective.

We have also observed that as clinical fidelity of a given scenario increases, so too does student criticism of minor inconsistencies between the scenario and their own unique work environments. This pattern of student reaction to clinical fidelity in training scenarios resembles the "uncanny valley" phenomenon [14], in which greater fidelity can be associated with increased criticism of observable discrepancies under certain conditions.

One potential solution to our scenario development process was to de-emphasize the clinical nature of scenarios in favor of a stronger non-clinical focus by expanding the training roles of our standardized patients.

2.0 INFRASTRUCTURE EVOLUTION

In order to maximize TeamSTEPPS training efficacy, a number of modifications and upgrades to the system's infrastructure were required. First, the existing audio/video system was upgraded to support team training. Second, the traditional model for healthcare simulation training scenarios was modified to de-emphasize clinical detail and focus on the non-clinical context of team training content.

2.1 Audio/Video System Upgrades

To enhance data collection and training debriefs, all ceiling cameras and microphones in the Simulation Center were upgraded to high-quality resolution systems. Additionally, a number of wall-mounted cameras were installed for the purpose of capturing team performance using eye-level panning. The wall-mounted cameras provide screen coverage of team performance otherwise unattainable by ceiling-mount cameras. The visualization control center was upgraded to include new monitors and selector switches for improving coordination among the cameras and microphones. A shoulder camera was purchased and incorporated into the data collection network, and a new dedicated computer system was installed for rendering hard-copy discs in minutes rather than hours (allowing for immediate debriefs).

2.2 Training Scenario Modifications

We expanded the utilization of our standardized patients by devising a new form of simulation training scenario. Rather than focusing on increased clinical fidelity for patient-centered, student-driven scenarios, we decided to pilot a series of scenarios which de-emphasize clinical details and focus instead on providing a high-fidelity social context in which students can practice non-clinical TeamSTEPPS skills. Our new scenarios each constitute a carefully scripted sequence of events which unfold in a generic healthcare setting, but do not involve clinical activities and are not patient-centric. This ensures that any student, regardless of background or job role, may freely participate in any of our scenarios.

The scripted scenario is carried out by a team of trained actors while students observe nearby. The team of actors engages in a sequence of social interactions with each other while the scenario unfolds, and some of the interactions are scripted to reflect sub-optimal teamwork decisions and behaviors. At various points, individual students are asked to step into the scenario as a participant and attempt to successfully resolve an escalating situation by drawing on their TeamSTEPPS training.

Each student is given multiple opportunities to engage the actors throughout the scenario. As they do so, they will receive realistic, immediate feedback from the actors in the form of improvised reactions. For example, a student who attempts to address an actor-physician's unprofessional behavior may receive a passive, hostile, or defensive response from the actor. This "interactive theater" simulation provides multiple opportunities for students to practice teamwork skills throughout and also supports continual student discussion/debriefing as part of the learning exercise.

3.0 IMPACT OF SYSTEMS EVOLUTION

3.1 Audio/Video Capabilities

As a result of the Simulation Center audio/video system upgrades, we are now able to record complete scenario sessions from multiple viewpoints and perspectives. We are also able to capture the entire student group in a single frame and identify sources of communication (including non-verbal) as events unfold. Complete audio/video integration and hard-copy disc transfer is possible within a matter of minutes, which permits relatively immediate video debriefs for students. This not only improves the quality of the training experience for learners, but also provides administrators with a record of how well the scenario functioned as a learning exercise. Additionally, recorded scenarios serve as training material for volunteer unit observers learning how to use TeamSTEPPS behavioral observation tools in the hospital.

3.2 Actor-driven Scenarios

Two non-clinical, actor-driven simulation scenarios were piloted in July 2010 at NCMP. Overall, the new format for TeamSTEPPS simulation training was considered a success. Student reactions to the actor-driven scenarios were positive. Because the emphasis was placed on social rather than clinical events, each scenario provided students multiple opportunities to engage without requiring a specific degree of clinical training or job role. This added flexibility gives us the ability to include students from a variety of backgrounds, including hospital administrative staff without any clinical training.

The new format also resulted in a greater amount of TeamSTEPPS-related dialogue during post-scenario debriefs, whereas clinical scenarios tend to be dominated by discussion of clinical activity and performance. Actor-driven event scripts guaranteed that the scenario would unfold in a manner consistent with our established learning objectives, whereas previous student-driven scenarios required constant interjection and management from staff scenario "directors."

4.0 FUTURE DIRECTION/GOALS

The NMCP Team Resource Center has adopted a theoretical framework for evaluating healthcare teams training efficacy based on Kirkpatrick's model of training outcomes assessment [1], and is using this framework to guide a systematic reconfiguration of the infrastructure supporting training and research initiatives at NMCP.

Early reaction-level feedback suggested that as efforts were made to increase the clinical fidelity of training scenarios, students were becoming more critical of minor inconsistencies within the clinical context. It was not our intention to emphasize technical proficiency with these scenarios, yet providing a practice environment with high clinical fidelity resulted in a preoccupation with clinical performance by our students. Therefore, the administrative team found it necessary to step back and reconsider the methodologies being used to develop TeamSTEPPS training scenarios.

Drawing on Benner's stages of clinical competence [15], our team began to reassess the students' readiness and progression with regard to TeamSTEPPS skill development. Benner's theory is based on the Dreyfus model of skill acquisition [16], which delineates five stages of increasing skill: novice, advanced beginner, competent, proficient, and expert. Our students were considered to be clinically proficient (to expert) within their own respective disciplines, yet were advanced beginners at best within the areas of communication and teamwork. Our primary goal was to facilitate student transformation from an advanced beginner in TeamSTEPPS to functional competence by the end of the 2.5-day training course.

Specifically, the identified learning objectives for course completion were a.) to demonstrate competence in the use of TeamSTEPPS strategies and techniques, b.) to be able to initiate TeamSTEPPS activities upon returning to the students' parent command, and c.) to recognize that developing the skills required to become a proficient TeamSTEPPS practitioner would require continued use of the strategies and techniques learned during the course. The distinction between the advanced beginner and competent skill levels was the guiding force behind our shift to a training scenario model emphasizing the social rather than clinical context.

The TRC's new model of actor-driven training scenarios reflects efforts to help students achieve TeamSTEPPS competence and to capitalize on Kirkpatrick's Level 3 (Behavior) training outcome [1]. The goal was to provide students with ample opportunities to apply TeamSTEPPS skills and strategies in a safe educational environment where immediate feedback could facilitate learning. Standardized patients are capable of providing students with two forms of feedback during these training scenarios: real-time improvisational feedback and post-scenario debrief feedback. The former constitutes a variety of realistic actor responses directed toward the students as they practice resolving teamwork issues throughout each scenario. The latter is an overall performance critique presented by the actor after the scenario has ended.

Standardized patient actors have been shown to be a reliable and valid means of assessing healthcare professionals' non-technical skills [17], [18]. The TRC is currently developing a standardized protocol for assessing students' TeamSTEPPS performance during simulated scenarios; the results of these assessments will serve as discussion during post-scenario debrief sessions. However, as with any formal assessment protocol, it will be critical to ensure that our assessments are not influenced by evaluator bias.

As we develop a standardized protocol for TeamSTEPPS skills assessment, we will examine the degree to which evaluator bias impacts ratings of student performance [18], [19]. Inconsistencies in actors' role portrayal, improvised feedback, or scoring could be the result of unique biases (e.g., gender, age) attributable to the actor-evaluators. One methodology that has been developed to assess standardized patient bias and establish inter-rater reliability is the use of "standardized examinees" [19]. Standardized examinees are individuals trained to a specific level of proficiency, after which they are subjected to assessment by a number of standardized patients. Inter-rater reliability can then be established and potential biases explored through the analysis of ratings provided by the various standardized patients.

The most important goals of the TRC moving forward are to continue stripping away the veneer of expertise that comes from students' confidence in their respective technical abilities and to reinforce the notion that a significant amount of learning still lies ahead for those who would develop teamwork expertise as well. It is our objective to provide meaningful learning experiences for students so that they complete the TeamSTEPPS course with the competence to implement TeamSTEPPS strategies and the motivation required to transform their competence into expertise.

5.0 REFERENCES

[1] Kirkpatrick, D.L. (1994). *Evaluating Training Programs: The Four Levels.* San Francisco, CA: Berrett-Koehler.

[2] Joint Commission on Accreditation of Healthcare Organizations (2006). *Root Causes for Sentinel Events.* Available at: http://www.jointcommission.org/SentinelEvents/Statistics/.

[3] Sorbero, M.E., Farley, D.O., Mattke, S., Lovejoy, S. (2008). Outcome Measures for Effective Teamwork in Inpatient Care (RAND technical report TR-462-AHRQ). Arlington, VA: RAND Corporation.

[4] Salas, E., Wilson, K.A., Burke, C.S., & Wightman, D.C. (2006). Does Crew Resource Management Training Work? An Update, an Extension, and Some Critical Needs. *Human Factors, 48*(2), 392-412.

[5] Salas, E., DiazGranados, D., Weaver, S.J., & King, H. (2008). Does Team Training Work? Principles for Health Care. *Academic Emergency Medicine, 15*, 1002-1009.

[6] Sexton, J.B., Thomas, E.J., & Helmreich, R.L. (2000). Error, Stress, and Teamwork in Medicine and Aviation: Cross Sectional Surveys. *British Medical Journal, 320*, 745-749.

[7] Agency for Healthcare Research and Quality (2006). *TeamSTEPPS: Team Strategies and Tools to Enhance Performance and Patient Safety Instructor Guide.* Washington, DC:

Agency for Healthcare Research and Quality.

[8] Nielsen, P.E. et al. (2007). Effects of Teamwork Training on Adverse Outcomes and Process of Care in Labor and Delivery. *Obstetrical & Gynecological Survey, 62*(5), 294-295.

[9] Moorthy, K., Munz, Y., Adams, S., Pandey, V., & Darzi, A. (2005). A Human Factors Analysis of Technical and Team Skills among Surgical Students during Procedural Simulations in a Simulated Operating Theatre. *Annals of Surgery, 242*(5), 631-639.

[10] Moorthy, K., Munz, Y., Forrest, D., Pandey, V., Undre, S., Vincent, C., & Darzi, A. (2006). Surgical Crisis Management Skills Training and Assessment: A Stimulation-Based Approach to Enhancing Operating Room Performance. *Annals of Surgery, 244*(1), 139-147.

[11] Berkenstadt, H., Haviv, Y., Tuval, A., Shemesh, Y., Megrill, A., Perry, A., Rubin, O., & Ziv, A., (2008). Improving Handoff Communications in Critical Care: Utilizing Simulation-based Training toward Process Improvement in Managing Patient Risk. *Chest, 134*(1), 158-162.

[12] Voss, J.D., May, N.B., Schorling, J.B., Lyman, J.A., Schectman, J.M., Wolf, A., Nadkarni, M.M., & Plews-Ogan, M. (2008). Changing Conversations: Teaching Safety and Quality in Residency Training. *Academic Medicine 83*(11), 1080-1087.

[13] Baker, D.P., Krokos, K.J., & Amodeo, A.M. (2008). *TeamSTEPPS teamwork attitudes questionnaire manual*. Washington, DC: American Institutes for Research.

[14] Mori, M. (1970). Bukimi No Tani (The Uncanny Valley). *Energy, 7*(4), 33-35.

[15] Benner, P. (2004). Using the Dreyfus model of skill acquisition to describe and interpret skill acquisition and clinical judgment in nursing practice and education. *Bulletin of Science, Technology & Society, 24*(3), 188-207.

[16] Dreyfus, S.E. & Dreyfus, H.L (1980). *A five-stage model of the mental activities involved in directed skill acquisition*. Washington, DC: Storming Media.

[17] McGovern, M.M., Johnston, M., Brown, K., Zinberg, R., & Cohen, D. (2006). Use of standardized patients in undergraduate medical genetics education. *Teaching and Learning in Medicine, 18*(3), 203-207.

[18] Van Zanten, M., Boulet, J.R., & McKinley, D. (2007). Using standardized patients to assess the interpersonal skills of physicians: Six years' experience with a high-stakes certification examination. *Health Communication, 22*(3), 195-205.

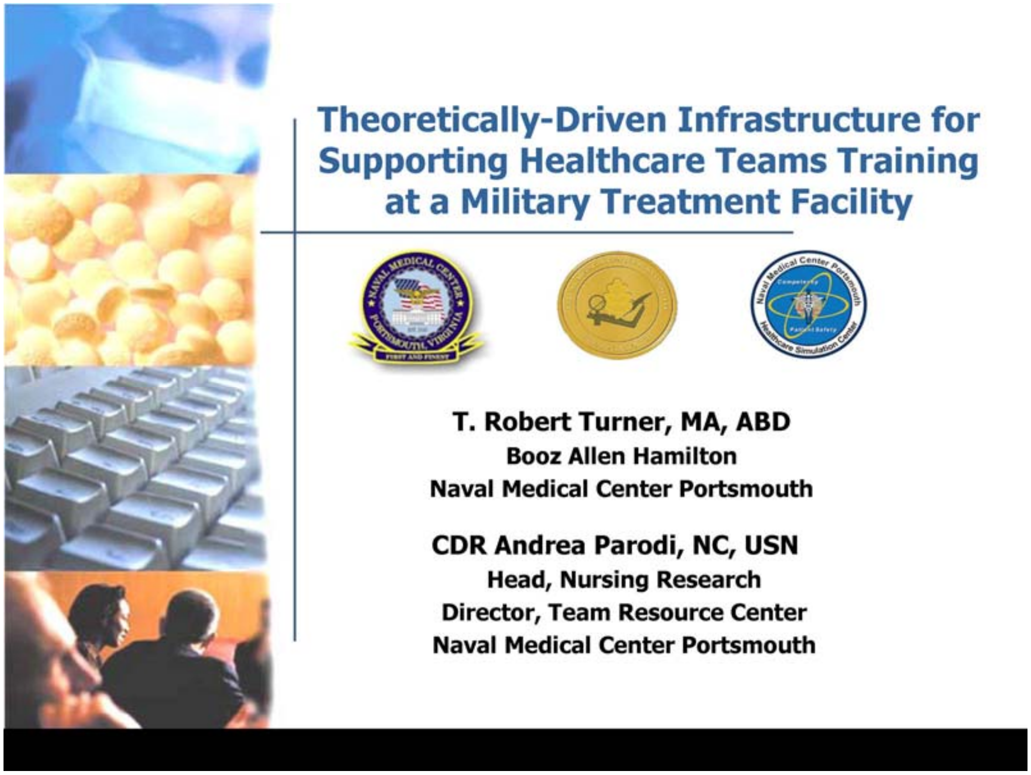
[19] Worth-Dickstein, H., Pangaro, L.N., MacMillan, M.K., Klass, D.J., & Shatzer, J.H. (2005). Use of "standardized examinees" to screen for standardized-patient scoring bias in a clinical skills examination. *Teaching and Learning in Medicine, 17*(1), 9-13.

6.0 ACKNOWLEDGMENTS

Support: TRICARE Management Activity, Office of the Secretary of Defense

Research data to be derived from *TeamSTEPPS Training Program Assessment Tool Revision and Validation*, an approved Naval Medical Center, Portsmouth, VA IRB/IACUC protocol (CIP# 2010.0015).

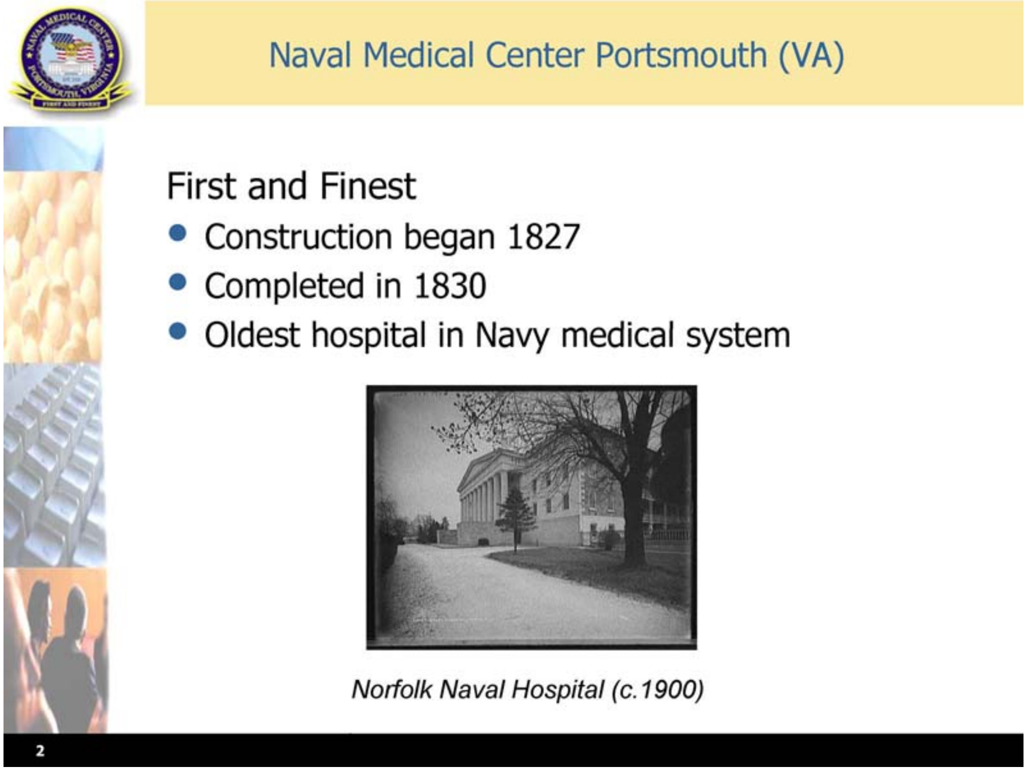
The views expressed in this paper are those of the author(s) and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense or the United States Government.



Theoretically-Driven Infrastructure for Supporting Healthcare Teams Training at a Military Treatment Facility

T. Robert Turner, MA, ABD
Booz Allen Hamilton
Naval Medical Center Portsmouth

CDR Andrea Parodi, NC, USN
Head, Nursing Research
Director, Team Resource Center
Naval Medical Center Portsmouth



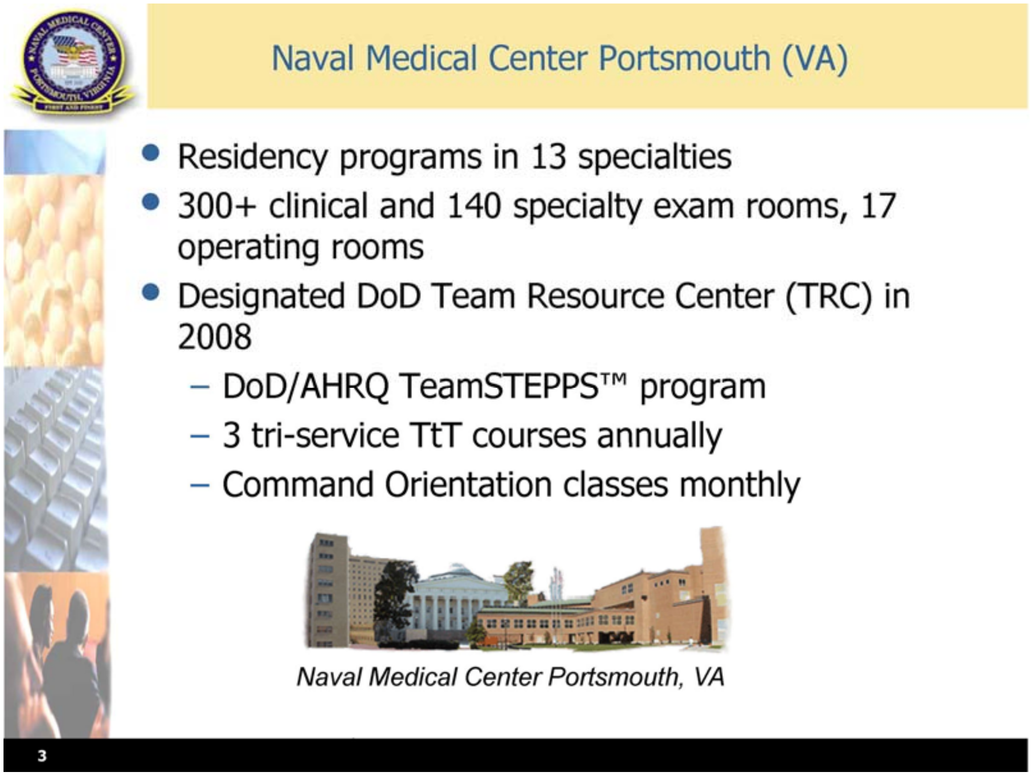
Naval Medical Center Portsmouth (VA)

First and Finest

- Construction began 1827
- Completed in 1830
- Oldest hospital in Navy medical system

Norfolk Naval Hospital (c.1900)

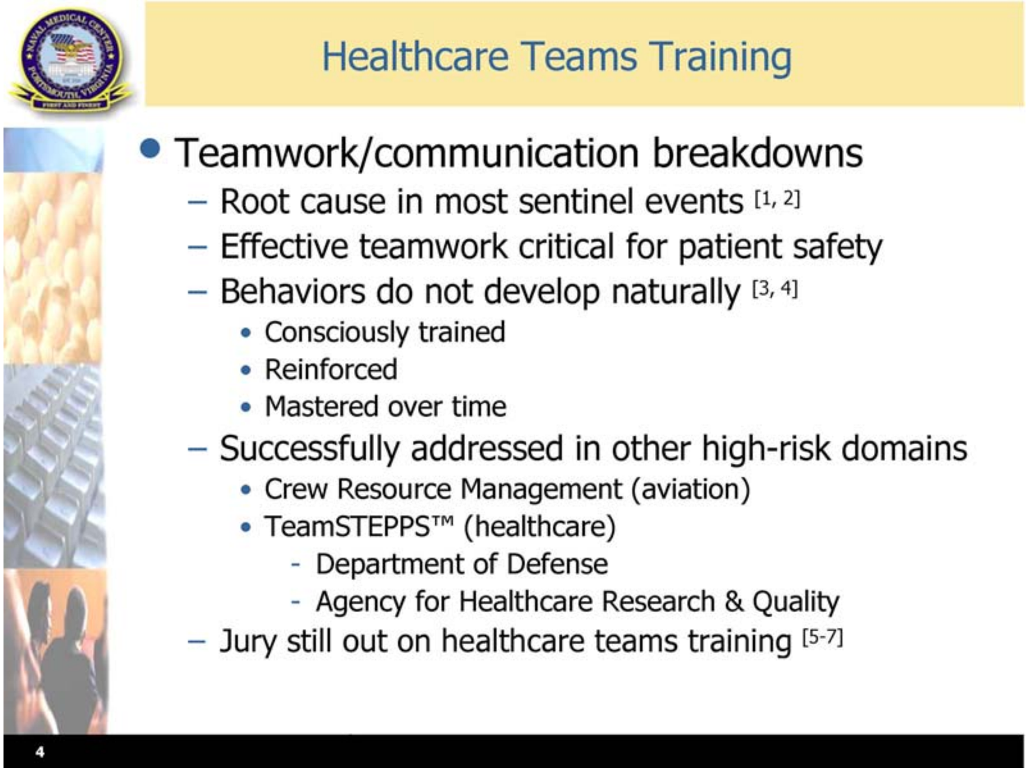
2



Naval Medical Center Portsmouth (VA)

- Residency programs in 13 specialties
- 300+ clinical and 140 specialty exam rooms, 17 operating rooms
- Designated DoD Team Resource Center (TRC) in 2008
  - DoD/AHRQ TeamSTEPPS™ program
  - 3 tri-service TtT courses annually
  - Command Orientation classes monthly

Naval Medical Center Portsmouth, VA



Healthcare Teams Training

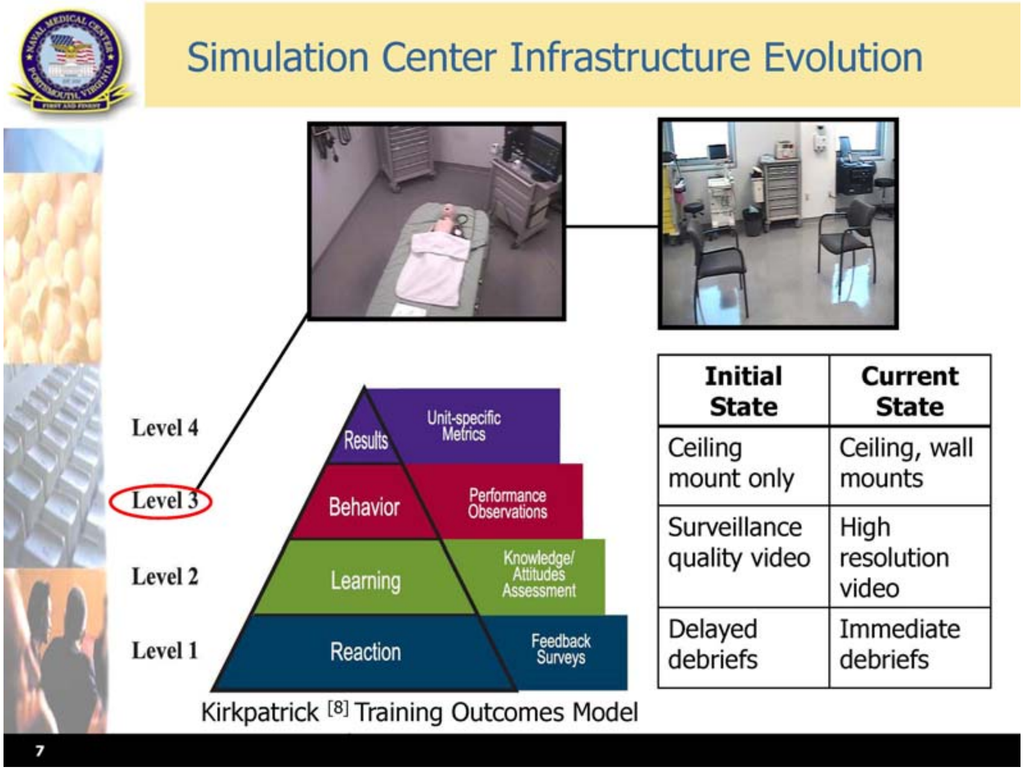
- Teamwork/communication breakdowns
  - Root cause in most sentinel events [1, 2]
  - Effective teamwork critical for patient safety
  - Behaviors do not develop naturally [3, 4]
    - Consciously trained
    - Reinforced
    - Mastered over time
  - Successfully addressed in other high-risk domains
    - Crew Resource Management (aviation)
    - TeamSTEPPS™ (healthcare)
      - Department of Defense
      - Agency for Healthcare Research & Quality
  - Jury still out on healthcare teams training [5-7]

TeamSTEPPS from CRM

- TeamSTEPPS core competencies
  - Team Structure
  - Leadership
  - Situation Monitoring
  - Mutual Support
  - Communication
- TRC program
  - Prepare to lead by example
  - Support change teams
  - Study modeling & simulation technologies

Healthcare Simulation Center

- Training/Research Assets
  - Didactic training facilities
  - Audio/Video capture system
  - Manikin
  - Virtual Reality
  - Standardized Patients

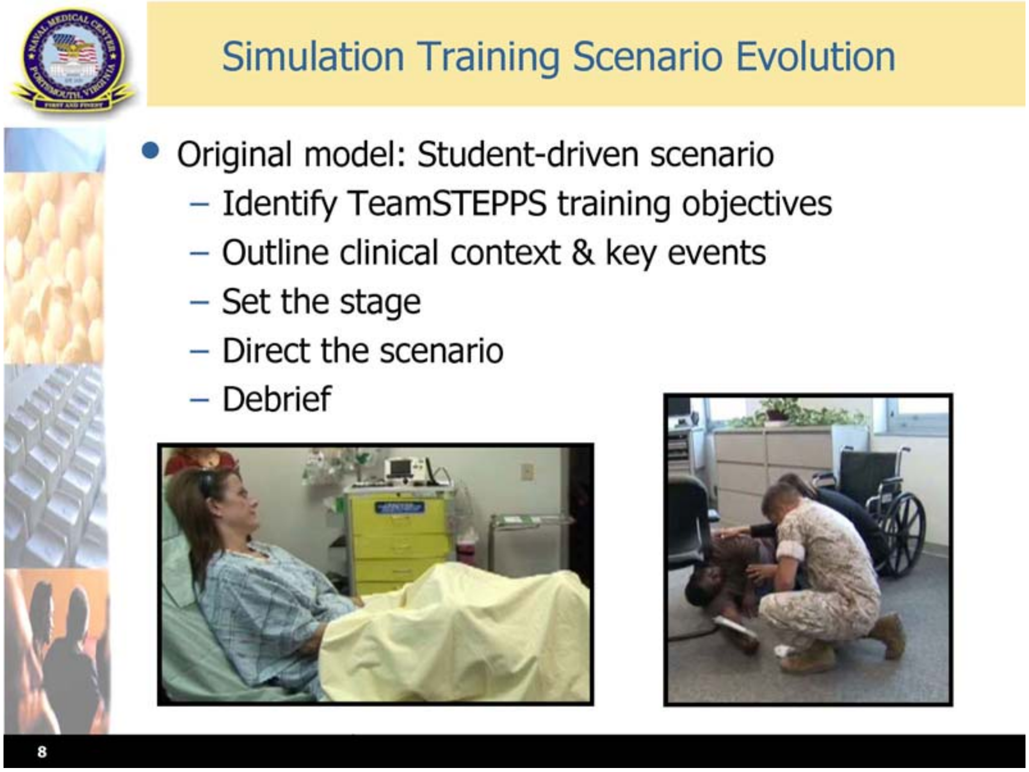


Simulation Center Infrastructure Evolution

Kirkpatrick [8] Training Outcomes Model

Initial State	Current State
Ceiling mount only	Ceiling, wall mounts
Surveillance quality video	High resolution video
Delayed debriefs	Immediate debriefs

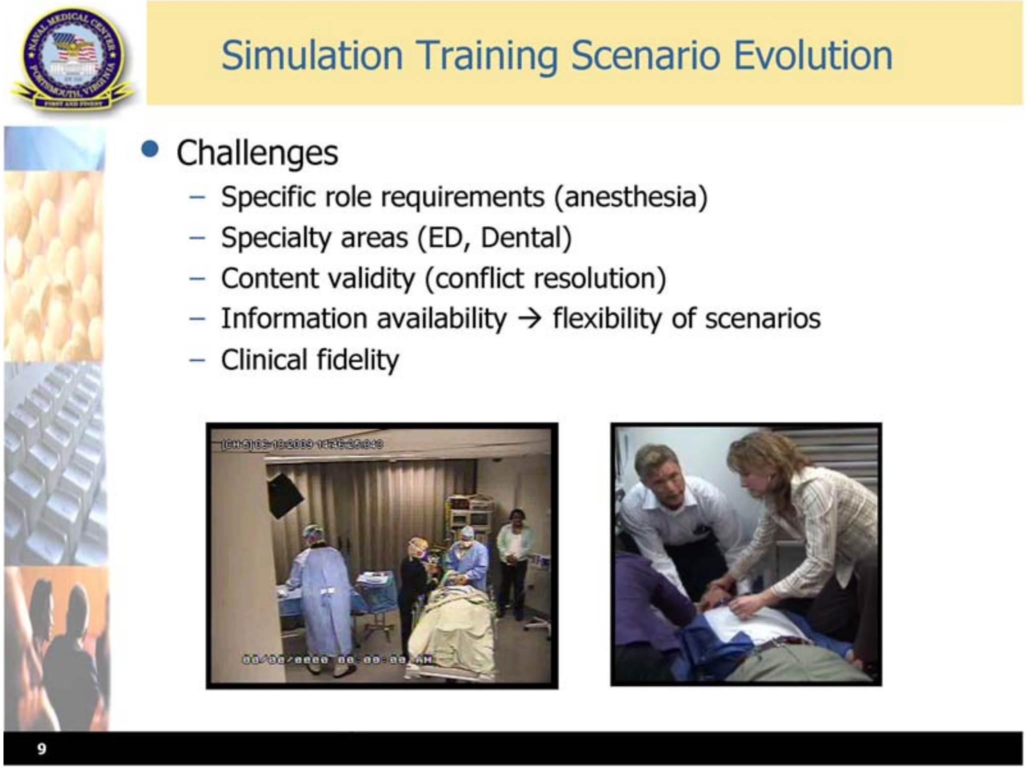
7



Simulation Training Scenario Evolution

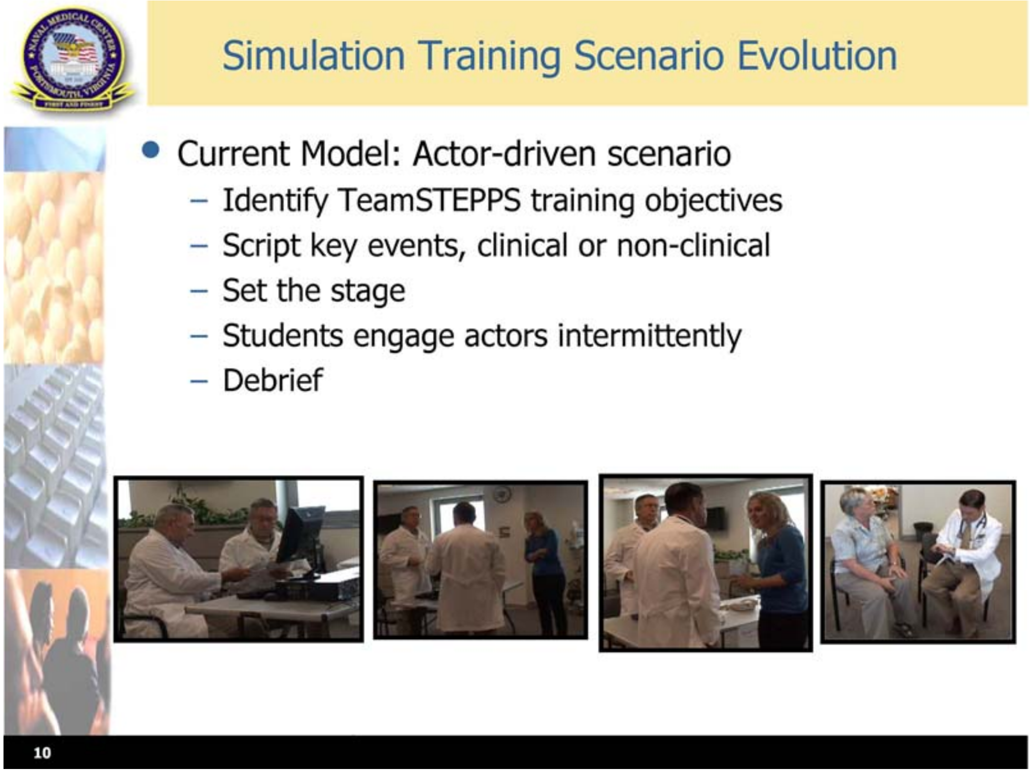
- Original model: Student-driven scenario
  - Identify TeamSTEPPS training objectives
  - Outline clinical context & key events
  - Set the stage
  - Direct the scenario
  - Debrief

8



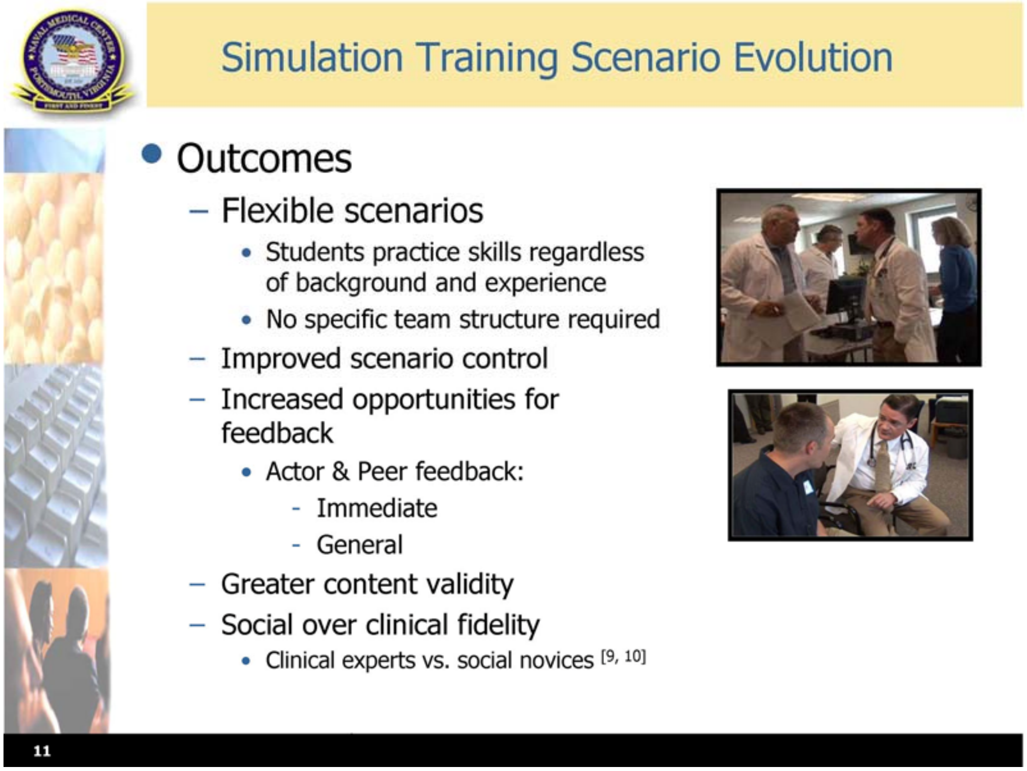
Simulation Training Scenario Evolution

- Challenges
  - Specific role requirements (anesthesia)
  - Specialty areas (ED, Dental)
  - Content validity (conflict resolution)
  - Information availability → flexibility of scenarios
  - Clinical fidelity



Simulation Training Scenario Evolution

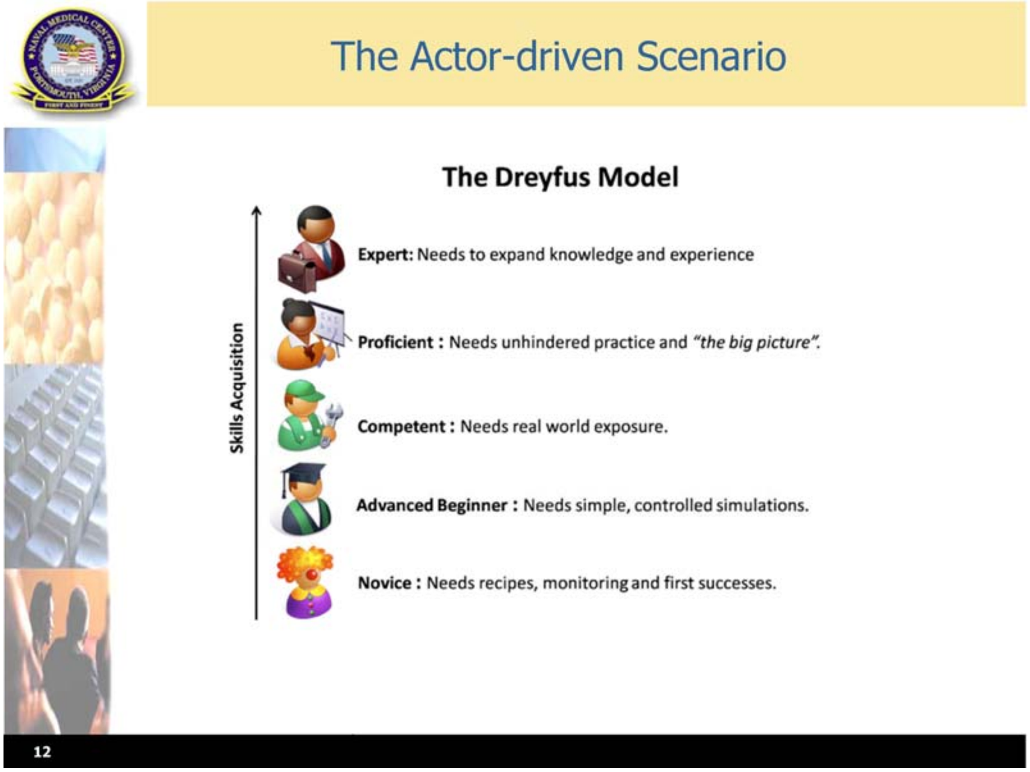
- Current Model: Actor-driven scenario
  - Identify TeamSTEPPS training objectives
  - Script key events, clinical or non-clinical
  - Set the stage
  - Students engage actors intermittently
  - Debrief



Simulation Training Scenario Evolution

- Outcomes
  - Flexible scenarios
    - Students practice skills regardless of background and experience
    - No specific team structure required
  - Improved scenario control
  - Increased opportunities for feedback
    - Actor & Peer feedback:
      - Immediate
      - General
  - Greater content validity
  - Social over clinical fidelity
    - Clinical experts vs. social novices [9, 10]

11



The Actor-driven Scenario

The Dreyfus Model

Expert: Needs to expand knowledge and experience

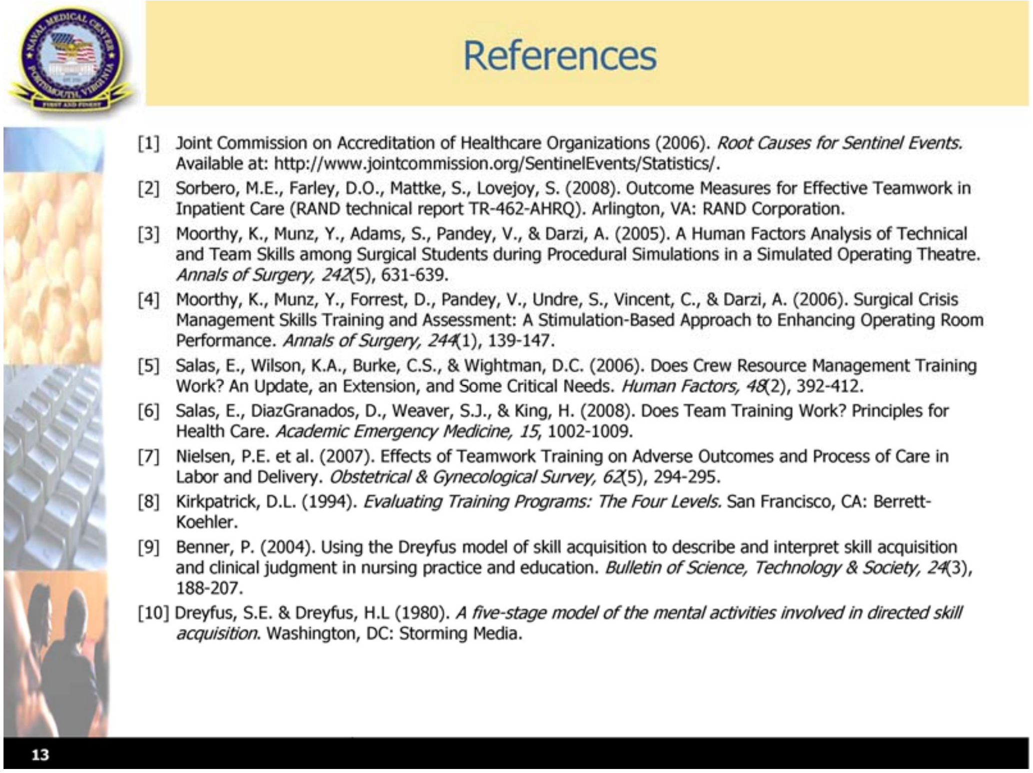
Proficient : Needs unhindered practice and *"the big picture".*

Competent : Needs real world exposure.

Advanced Beginner : Needs simple, controlled simulations.

Novice : Needs recipes, monitoring and first successes.

12



References

[1] Joint Commission on Accreditation of Healthcare Organizations (2006). *Root Causes for Sentinel Events.* Available at: http://www.jointcommission.org/SentinelEvents/Statistics/.

[2] Sorbero, M.E., Farley, D.O., Mattke, S., Lovejoy, S. (2008). Outcome Measures for Effective Teamwork in Inpatient Care (RAND technical report TR-462-AHRQ). Arlington, VA: RAND Corporation.

[3] Moorthy, K., Munz, Y., Adams, S., Pandey, V., & Darzi, A. (2005). A Human Factors Analysis of Technical and Team Skills among Surgical Students during Procedural Simulations in a Simulated Operating Theatre. *Annals of Surgery, 242*(5), 631-639.

[4] Moorthy, K., Munz, Y., Forrest, D., Pandey, V., Undre, S., Vincent, C., & Darzi, A. (2006). Surgical Crisis Management Skills Training and Assessment: A Stimulation-Based Approach to Enhancing Operating Room Performance. *Annals of Surgery, 244*(1), 139-147.

[5] Salas, E., Wilson, K.A., Burke, C.S., & Wightman, D.C. (2006). Does Crew Resource Management Training Work? An Update, an Extension, and Some Critical Needs. *Human Factors, 48*(2), 392-412.

[6] Salas, E., DiazGranados, D., Weaver, S.J., & King, H. (2008). Does Team Training Work? Principles for Health Care. *Academic Emergency Medicine, 15*, 1002-1009.

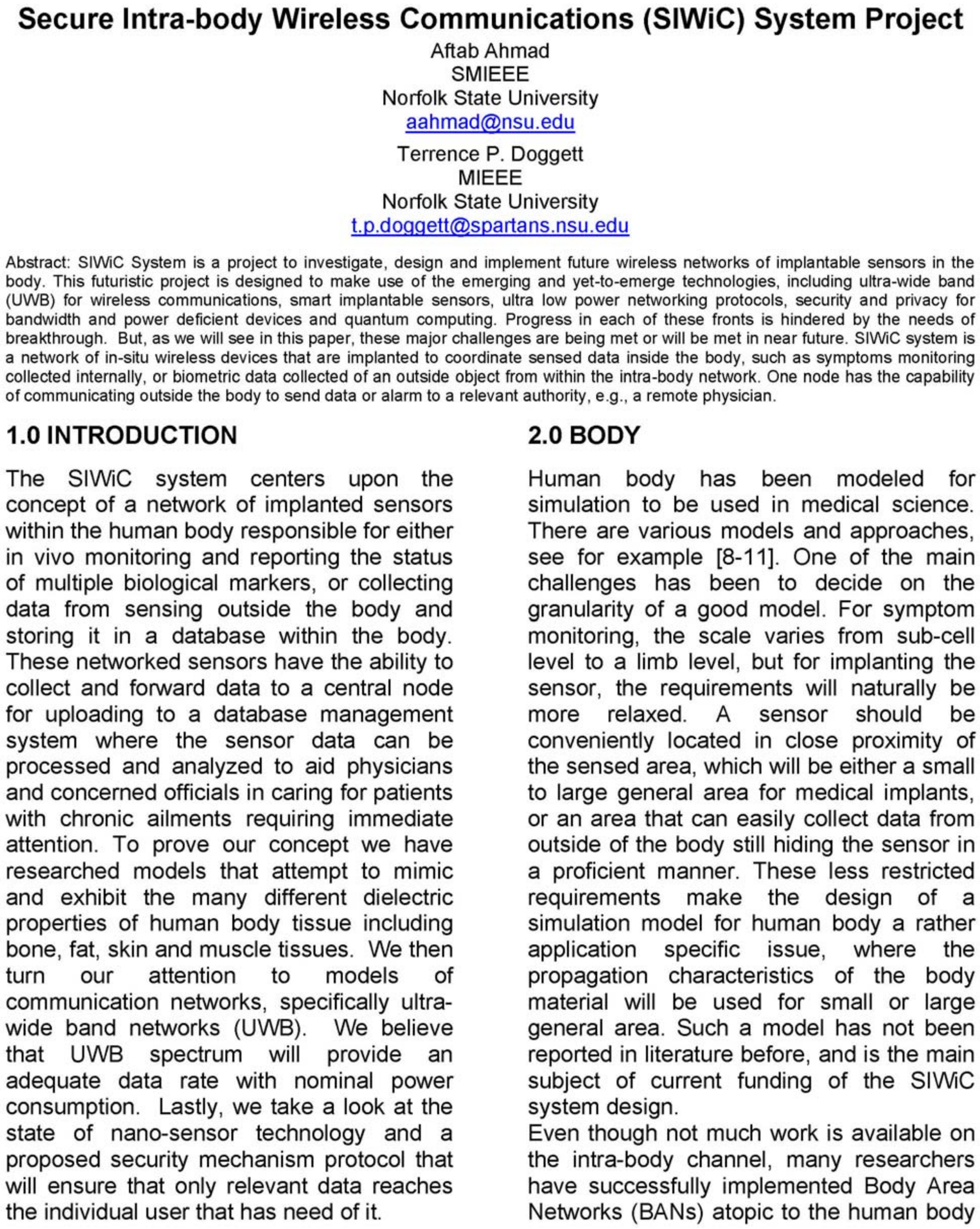
[7] Nielsen, P.E. et al. (2007). Effects of Teamwork Training on Adverse Outcomes and Process of Care in Labor and Delivery. *Obstetrical & Gynecological Survey, 62*(5), 294-295.

[8] Kirkpatrick, D.L. (1994). *Evaluating Training Programs: The Four Levels.* San Francisco, CA: Berrett-Koehler.

[9] Benner, P. (2004). Using the Dreyfus model of skill acquisition to describe and interpret skill acquisition and clinical judgment in nursing practice and education. *Bulletin of Science, Technology & Society, 24*(3), 188-207.

[10] Dreyfus, S.E. & Dreyfus, H.L (1980). *A five-stage model of the mental activities involved in directed skill acquisition*. Washington, DC: Storming Media.

3.5 Secure Intra-body Wireless Communications (SIWiC) System Project



Secure Intra-body Wireless Communications (SIWiC) System Project

Aftab Ahmad
SMIEEE
Norfolk State University
aahmad@nsu.edu

Terrence P. Doggett
MIEEE
Norfolk State University
t.p.doggett@spartans.nsu.edu

Abstract: SIWiC System is a project to investigate, design and implement future wireless networks of implantable sensors in the body. This futuristic project is designed to make use of the emerging and yet-to-emerge technologies, including ultra-wide band (UWB) for wireless communications, smart implantable sensors, ultra low power networking protocols, security and privacy for bandwidth and power deficient devices and quantum computing. Progress in each of these fronts is hindered by the needs of breakthrough. But, as we will see in this paper, these major challenges are being met or will be met in near future. SIWiC system is a network of in-situ wireless devices that are implanted to coordinate sensed data inside the body, such as symptoms monitoring collected internally, or biometric data collected of an outside object from within the intra-body network. One node has the capability of communicating outside the body to send data or alarm to a relevant authority, e.g., a remote physician.

1.0 INTRODUCTION

The SIWiC system centers upon the concept of a network of implanted sensors within the human body responsible for either in vivo monitoring and reporting the status of multiple biological markers, or collecting data from sensing outside the body and storing it in a database within the body. These networked sensors have the ability to collect and forward data to a central node for uploading to a database management system where the sensor data can be processed and analyzed to aid physicians and concerned officials in caring for patients with chronic ailments requiring immediate attention. To prove our concept we have researched models that attempt to mimic and exhibit the many different dielectric properties of human body tissue including bone, fat, skin and muscle tissues. We then turn our attention to models of communication networks, specifically ultra-wide band networks (UWB). We believe that UWB spectrum will provide an adequate data rate with nominal power consumption. Lastly, we take a look at the state of nano-sensor technology and a proposed security mechanism protocol that will ensure that only relevant data reaches the individual user that has need of it.

2.0 BODY

Human body has been modeled for simulation to be used in medical science. There are various models and approaches, see for example [8-11]. One of the main challenges has been to decide on the granularity of a good model. For symptom monitoring, the scale varies from sub-cell level to a limb level, but for implanting the sensor, the requirements will naturally be more relaxed. A sensor should be conveniently located in close proximity of the sensed area, which will be either a small to large general area for medical implants, or an area that can easily collect data from outside of the body still hiding the sensor in a proficient manner. These less restricted requirements make the design of a simulation model for human body a rather application specific issue, where the propagation characteristics of the body material will be used for small or large general area. Such a model has not been reported in literature before, and is the main subject of current funding of the SIWiC system design.

Even though not much work is available on the intra-body channel, many researchers have successfully implemented Body Area Networks (BANs) atopic to the human body in the form of wearable sensors [1] and still others have been successful with standalone, non-networked implanted devices that monitor a specific function of interest [2]. Of greater interest is a network of implanted devices that each monitor a specific function, are capable of exchanging data among themselves and with a central node. This central node will also offload data to a remote server for further processing and storage. Such a network will use human tissue as medium to propagate data signals from one node to another. In other words, the nodes in the network will use human bone, muscle and fat to convey signals through the body to other nodes in the much the same way as sound travels from a source through air or water to a receiver.

2.1 SIWiC Node

A typical SIWiC node consists of a sensing element, ideally a smart sensor such as conforming to IEEE specifications [27], that senses analog data and uses inductive coupling or direct connection to the sensed area. For inductive coupling, a transducer converts the sensed data on location into equivalent electrical signal and this signal is picked up by a coil at a distance (e.g., both being on different sides of a membrane). The signal, analog in nature, passes through an analog to digital convertor (ADC), which generates a digital stream to be packetized by a network protocol, such as IEEE 802.15.6 [28] or IEEE 802.15.4a [26] that sends packets to an UWB physical layer. In case of a smart sensor, the digitization occurs inside the sensor assembly and the output of the sensor is ready for UWB network. The output of UWB module is secured and networked in general to a database inside the body. The signal will be weak enough so that it does not travel outside the body, except for a node with the 'symptoms' database (SDB) that will be highly secured and linkable externally using a secure protocol. Figure 1 shows a schematic of the SIWiC node interconnection.

Most work in the direction of implementation for the SIWiC node is yet to be done. In this paper, we address the channel modeling for the networking part where human body acts as a wireless communications medium.

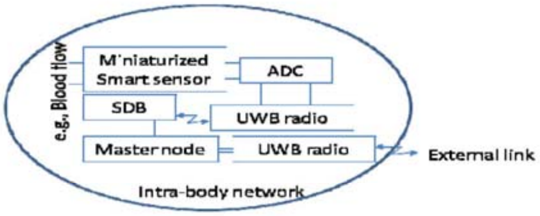


Figure 1. SIWiC System node and its relation to SDB.

2.2 Existing Efforts

In our research we have reviewed a few different ideas/models for implementing a Body Area Network and even some ideas on internal BAN channels. In this section we will expound on a few of the more promising approaches to include ideas from the Swiss Federal Institute of Technology-Zurich (ETH), the National Institute of Standards and Technology(NIST), the Electronics and Telecommunications Research Institute(ETRI) and Samsung's Electric Field Communication. We will also, very briefly, discuss some of the human dielectric properties that are important factors in modeling the human body as a communication channel.

First we will discuss the model proposed by the Swiss Institute of Technology-Zurich. In this paper galvanic coupling is presented as a possible transmission mechanism inside the human body. Galvanic coupling was investigated by Oberle [4] and analyzed by Hachisuka et al [5] [6]. Galvanic coupling uses the body's electrical potential stored in each cell to generate the electrical energy needed to propagate the data signal from source to receiver through cell clusters inside the human body. The signal is applied differentially over two transmitter electrodes and received differentially by two receiver electrodes [7]. The source electrical signal is modulated into the body and received differentially by the receiver electrodes as shown in Figure 2.

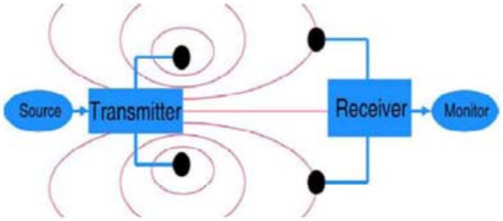


Fig 2. Signal Propagation via galvanic coupling IEEE Transactions on Biomedical Engineering, VOL 54, No. 10, October 2007

This method uses current with a peak amplitude of 1 mA between 10 KHz and 1MHz establishing a certain potential distribution in the human body. To determine paths of least resistance Electrical Impedance Tomography (EIT) and Finite Element (FE) modeling is used to map the impedance distribution of human tissues [8] – [11].

The IEEE P802.15 Working Group for Wireless Personal Area Networks (WPANs) is the next statistical model we will discuss. This model was developed within the National Institute for Standards and Technology and proposes a simple statistical model for representing the path loss for communication to/from an implant or between two implants inside a human body. The path loss model is modeled on the center frequency of 403.5 MHz, the Medical Implant Communication Service (MICS) band. This is an unlicensed band allocated for communication between an implanted medical device and an external controller, with these primary benefits; better propagation characteristics for implants, reasonable size antenna for implants, worldwide availability and limited threat of interference. It is our point of view that the required antenna size and power are still way too high and only UWB can satisfy these requirements at this time.

To study the propagation characteristics of MICS, NIST proposed a 3D simulation & visualization scheme because in-body measurement and experiments prove very difficult at this point in time. The 3D simulation & visualization system components consist of dielectric properties from over 300 male human body parts that are user definable. The system uses a 3D full-wave electromagnetic field simulation (HFSS) propagation engine accurate to within 2mm. Some of the user-selectable input parameters include antenna location, antenna orientation, operating frequency, transmit power, range and resolution. The study simulated near surface implants such as pacemakers and deep-tissue implants in the form of endoscopy capsules 95 mm below the body surface. The resulting data was filtered into 3 sets: in-body propagation, all points completely within the body; in-body to body-surface and in-body to out-body propagation. See Figure 3 for NIST model.

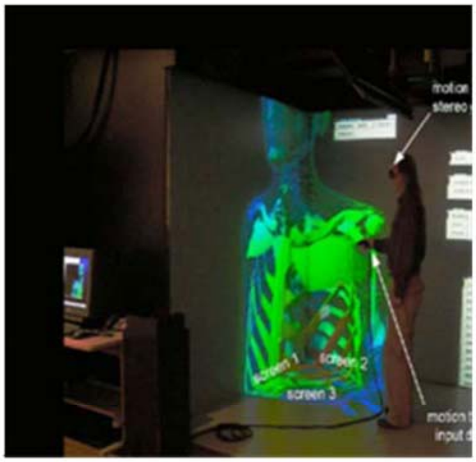


Fig 3. NIST 3D Simulation & Visualization Model, IEEE 802.15-08-0519-01-0006

The Electronics and Telecommunications Research Institute (ETRI) HBC (Human Body Communication) PHY Proposal for Body Area Network suggests that the physical channel of the body area network should have a data rate of 10 Kbps to 10 Mbps, be of low complexity, low power consumption and within regulatory compliance and operate from distances of 1 m to 3 m. The HBC should feature touch and play operation, be intuitive and easy to setup and use. Privacy and security should be afforded and power consumption should be extremely low.

The system principles include a frequency response range of 5 MHz ~ 50 MHz using a FSBT (Frequency Selective Baseband Transmission.) The FSBT allows a direct digital transmission without the need of a radio and offers more processing gain while avoiding low frequencies. FSBT employs sub-groups of Walsh codes and spreading techniques to increase processing gain. The HBC consists of a microcontroller, a signal electrode and the modem with its various sub-components (i.e. transmitter, receiver, buffers and noise filters, see Figure 4). The simulation parameters include a data rate of up to 10 Mbps, a maximum chip rate of 64 Mcps, the Walsh spreading code, and a baseband transmission square wave.

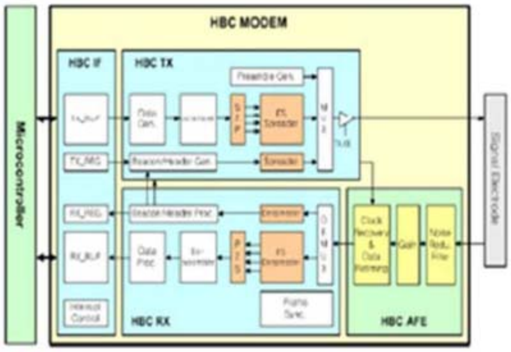


Fig 4. Human Body Communication System Architecture, IEEE 802.15-09-0348-00-0006

The Physical and MAC Layer proposal for the IEEE 802.15.6 Working Group for Wireless Personal Area Networks also received a model from Samsung Electronics. The proposal is for non-medical applications in an everyday environment such as entertainment or home office. The WPAN is intended for on-body to on-body communication with data rate ranges up to several Mbps. Lastly, as with all wireless personal area networks, low power consumption is critical. Communication is based on an Electric Field using the human body as the dielectric material, as shown in Figure 5. It has been determined that the human body has about 300-500 times more permittivity than air [12].

The Electric Field Communication physical layer consists of the transmitter which employs orthogonal modulation using frequency shaping coding, the sensor electrode, the medium (human body) the receiver electrode and the receiver which filters the data, compares it, and outputs the raw data. The data rate is scalable and the frequency bands range between 10 and 50 MHz. The packet structure of the PHY layer consists of the preamble, start frame delimiter and the payload. The payload maximum size is < 1 K Octets and is further divided into the MAC Header, MAC payload and the Frame Check Sequence. The receiver makes use of a single electrode instead of a 50 Ω antenna which eliminates the need for RF carrier signal blocks.

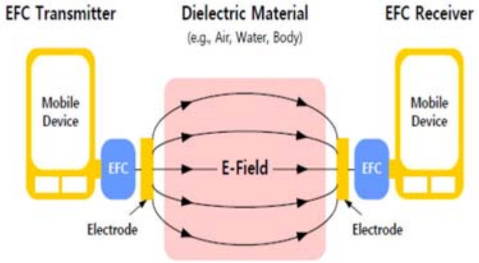


Fig 5. Electric Field Communication Architecture , IEEE 802.15-09-0318-02-006

2.3 UWB Models

In order to better understand PAN channel models we reviewed a number of models for Personal Area Networks. For example, the authors of [24] describe a few of the more popular UWB communication models and the primary characteristics of the multipath channels that make them up; root mean square (RMS) delay-spread, power decay profile and number of multipath components. Three indoor channel models were considered; the tap-delay line Rayleigh fading model [3], the Saleh-Valenzuela (S-V) model [4] and the Δ-R model described in [5]. Each model was suitably modified to fit the important channel characteristics described above. Of the three, the (S-V) model was chosen to model the multipath of an indoor environment for wideband channels on the order of 100 MHz. This model requires four main parameters to describe an environment, which can be adjusted for different environments; the cluster arrival rate, the ray arrival rate within a cluster, the cluster delay factor, and the ray decay factor.

In [13] an electromagnetic "creeping wave" is considered for modeling a communication channel around the human body. The authors have concluded that electromagnetic waves can travel through the human body or in a near surface channel around the body. This channel around the body is the one of interest and provides the least amount of delay and path loss, see Figure 6. To simulate this path around the human body a Finite-Difference Time-Domain EM simulator is used the XFDTD by REMCOM. Frequencies in the ISM bands (US) of 315 MHz, 915 MHz and 2.4 GHz are used because these are the bands that will most likely be used to develop a prototype and also because these higher bands allow the use of smaller antennas[13].

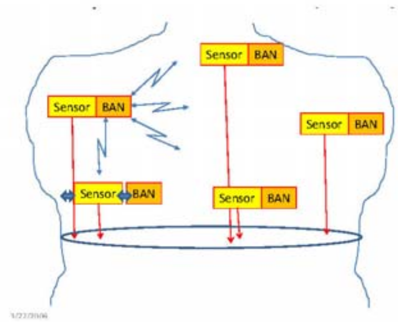


Fig 6. Wireless Communication Channel Around Human Body

Much work is needed in order to come up with a UWB model for human body that will predict the most realistic amount of transmission power needed to exchange a signal between two devices within the body.

3.0 DISCUSSION

3.1 State of Nanosensor Technology

There are many new companies using nanotechnology in very exciting ways in all industries across the board from military aircraft to cars, bicycles, and tennis rackets [25]. A new company known as NanoDynamics plans to use nanomaterials to bring about less weight shift inside of golf balls as they spin, drastically reducing hooks and slices [15]. Nanotechnology is a collection of technologies for building materials and devices "from the bottom up," atom by atom and usually includes items on the length scale of approximately 1-100 nanometers.[16] In our research we are primarily concerned with nanosensors; devices used to sense biological markers inside of the human body. Nano-sensors are already in the market [25] and are expected in near future to go to a scale where they can be implanted.

3.2 Security

It is well known in the health care field that patient privacy and rights are of utmost importance. SIWiC will use a novel wireless security protocol that satisfies the requirements of the HIPAA (Health Insurance Portability and Accountability Act) in the United States while maintaining timely access to vital patient information. HIPAA provides the guidelines for managing healthcare information and patient's privacy rights, ensuring that only individuals with a need-to-know have access to a patient's personal healthcare data.[19-22] Satisfying the requirements of HIPAA, in some instances, may prove to hinder the timely care of some critical patients and therefore must be addressed.

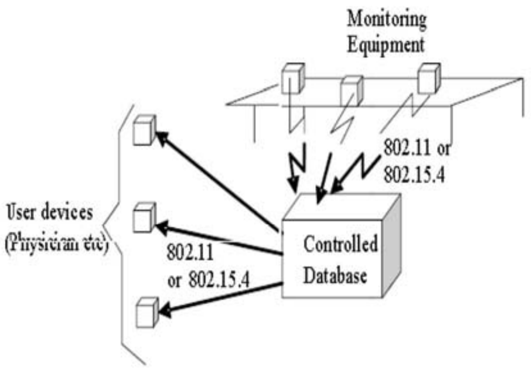


Figure 7. Privacy in a HIPAA compliance system can be implemented by sending all data to a controlled database.

A security protocol inside the symptoms database (SDB) addresses this issue by introducing a centralized database into the system architecture responsible for determining access levels based on a *HIPAA filter* and previously stored information tables.[23] Individuals and entities are assigned levels of access and stored in the database. Patient information is received at the database and immediately filtered to determine the type of information and assigned a classification level. Once a wireless request is received the system checks the credentials of the requestor against the level of information requested and determines whether access is granted or denied.

4.0 CONCLUSION

Given the current state of technology and the advances in nanosensor technology, a network of implantable sensors is feasible in near future, and will be a state-of-the-art in future. The hurdles are few and include adapting or developing an optimal human dielectric tissue model, adequately powering our sensor nodes and implantation of the nodes themselves. Additionally, although not addressed in this paper, the social stigma of having monitoring devices implanted within the body, by most individuals would have to be overcome. For the patients that stand to gain, by living fuller, longer and healthier lives, the benefits far outweigh the drawbacks associated with Implantable Sensor Networks.

5.0 REFERENCES

[1] A.T. Barth, M.A. Hanson, H.C. Powell, Jr., D. Unluer, S.G. Wilson, J. Lach, "Body-Coupled Communication for Body Sensor Networks", ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) Artticle No. 12, 2008.

[2] K. Gaffney, "Implanted Medical Device Aims to Lower Blood Pressure", University of Rochester Medical Center, http://www.medicalnewstoday.com/articles/22140.php , 03 Apr 2005.

[3] D. Halperin, T.S. Heydt-Benjamin, K. Fu, T. Kohno, W.H.Maisel, "Security and Privacy for Implantable Medical Devices", Pervasive Computing, IEEE Computer Society, Vol. 7, No.1 January-March 2008.

[4] M. Oberle, "Low power system-on-chip for biomedical application," Ph.D. dissertation, Integrated Syst. Lab. (IIS), ETH Zurich, Zurich, Switzerland, 2002.

[5] K. Hachisuka, A. Nakata, T. Takeda, Y. Terauchi, K. Shiba, K. Sasaki, H. Hosaka, and K. Itao, "Development and performance analysis of an intra-body communication device," in Proc. 12th Int. Conf. Solid State Sensors, Actuators Microsyst. (Transducers), 2003, vol. 2, pp. 1722–1725.

[6] K. Hachisuka, Y. Terauchi, Y. Kishi, T. Hirota, K. Sasaki, H. Hosaka, and K. Ito, "Simplified circuit modeling and fabrication of intra-body communication devices," in Proc. 13th Int. Conf. Solid-State Sensors, Actuators Microsyst., 2005, vol. 2E4-3, pp. 461–464.

[7] M.S. Wegmueller, A. Kuhn, J.Froehlich, M. Oberle "An Attempt to Model the Human Body as a Communication Channel", IEEE Transactions on Biomedical Engineering, VOL 54, NO. October 2007

[8] D. Barber and B. Brown, "Applied Potential Tomography," J. Phys. E,Sci. Instrum., vol. 17, pp. 723–733, 1984.

[9] N. Polydorides and W. R. B. Lionheart, "A MATLAB toolkit for three-dimensional electrical impedance tomography: A contribution to the electrical impedance and diffuse optical reconstruction software project," Meas. Sci. Technol., vol. 13, no. 12, pp. 1871–1883, 2002.

[10] G. J. Saulnier, R. S. Blue, J. C. Newell, D. Isaacson, and P. Edic, "Electrical Impedance Tomography," IEEE Signal Process. Mag., vol. 18, no. 6, pp. 31–43, Nov. 2001.

[11] A. V. Shahidi, R. Guardo, and P. Savard, "Impedance tomography—Computational Analysis based on Finite-Element Models of a Cylinder and a Human Thorax," Ann. Biomed. Eng., vol. 23, no. 1, pp. 61–69, 1995.

[12] J.S. Park, et al, "Samsung EFC PHY & MAC Proposal", IEEE 802.15-09-0318-02-0006, May 2009.

[13] J. Ryckaert, P. De Doncker, R. Meys, A. de Le Hoye and S. Donnay, "Channel Model for Wireless Communication Around Human Body", Electronics Letters 29th April 2004, Vol. 40, No. 9.

[14] M. Jacoby, "Composite Materials: Custom Blending of Materials with Distinct Characteristics Leads to Advanced Composites with Tailor-made Properties", CHEMICAL & ENGrNEERiNG NEWS, Aug. 30, 2004, at 34; Tom Henderson, Nanotech May Help Autos Cut Fuel Use, DETROIT NEWS, NOV. 28,2004, at C3.

[15]K. Maney, "Nanotech Could Put a New Spin on Sports: One Example Golf Balls That Make Hacks Look Good", USA TODAY, NOV. 17, 2004, at 1B.

[16]ENVTL. PROT.AGENCY,NANOTECHNOLOGY:BASICINFORMATION, http://es.epa.gov/ncer/nano/questions/.

[17]Business Communications Company, "Nanotechnology for Photonics", January 2005.

[18]Ainsworth, *supra* note 80, at 19, *See also* Vivek Koppikar et al,, "Current Trends in Nanotech Patents: A View from Inside the Patent Office", 1 NANOTECHNOLOGY L, & Bus, J, 24 (2004),

[19] "HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT OF 1996," Public Law 104-191,104thCongress, http://aspe.hhs.gov/admnsimp/pl104191.htm, 1996.

[20] Office for Civil Rights, "Summary of HIPAA Privacy Rule," United States Department of Health and Human Services, 2003.

[21] K.Beaver and R. Herold, "The Practical Guide to HIPAA Privacy and Security Compliance", Auerbach, 2003.

[22] T. Jepsen, "IT in Healthcare: Progress Report," IT Professional, vol. 5, no. 1, pp. 8-14, 2003.

[23] A. Ahmad, A. Riedl, W.J. Naramore, Nee-Yin, Chon & M. Alley, "Comparative Study of Security in IEEE 802.11-2007 and IEEE 802.15.4-2006 for Patient Monitoring Environments", Jan 2010.

[24] A.F. Molisch, "Channel Models for UltraWideBand Personal Area Networks", Mitsubishi Electric Research Labs, IEEE 1536-1284/03, Dec 2003.

[25] M.A. Van Lente, "Building the New World of Nanotechnology", Case Western Reserve Journal of International Law, Vol. 38:173, 2006.

[26] "IEEE Standard for Information technology-Telecommunications and information exchange between systems- Local and metropolitan area networks- Specific requirements Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low- Rate Wireless Personal Area Networks (WPANs)," IEEE Std 802.15.4-2006 (Revision of IEEE Std 802.15.4-2003) , pp. 1-305, 2006.

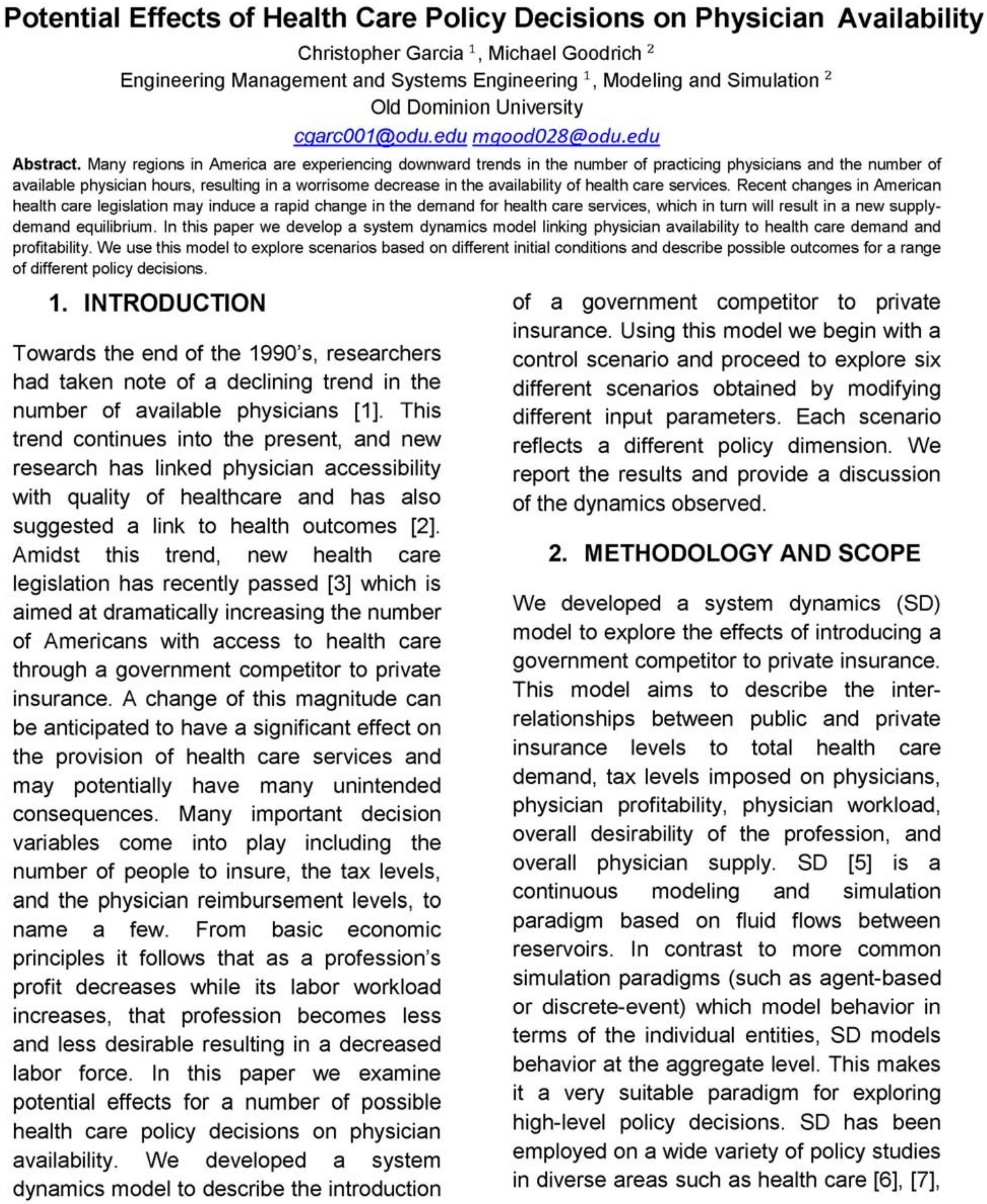
[27] "IEEE Standard for a family of Smart Transducer Interface Standards, describes a set of open, common, network-independent communication interfaces for connecting transducers (sensors or actuators) to microprocessors, instrumentation systems, and control/field networks.", IEEE 1451 FAMILY OF SMART TRANSDUCER INTERFACE STANDARDS, http://standards.ieee.org/regauth/1451/index.html

[28] "The IEEE 802.15 Task Group 6 (BAN) is developing a communication standard optimized for low power devices and operation on, in or around the human body (but not limited to humans) to serve a variety of applications including medical, consumer electronics / personal entertainment and other.", http://www.ieee802.org/15/pub/TG6.html.

6.0 ACKNOWLEDGMENTS

The authors would like to acknowledge that some of the work for SIWiC System design is supported by National Reconnaissance Office (NRO) via a grant to the Norfolk State University on "Implantable Sensor Networks: Human Body as a Wireless Communications Medium".

3.6 Potential Effects of Health Care Policy Decisions on Physician Availability



Potential Effects of Health Care Policy Decisions on Physician Availability

Christopher Garcia [1], Michael Goodrich [2]
Engineering Management and Systems Engineering [1], Modeling and Simulation [2]
Old Dominion University
cgarc001@odu.edu mgood028@odu.edu

Abstract. Many regions in America are experiencing downward trends in the number of practicing physicians and the number of available physician hours, resulting in a worrisome decrease in the availability of health care services. Recent changes in American health care legislation may induce a rapid change in the demand for health care services, which in turn will result in a new supply-demand equilibrium. In this paper we develop a system dynamics model linking physician availability to health care demand and profitability. We use this model to explore scenarios based on different initial conditions and describe possible outcomes for a range of different policy decisions.

1. INTRODUCTION

Towards the end of the 1990's, researchers had taken note of a declining trend in the number of available physicians [1]. This trend continues into the present, and new research has linked physician accessibility with quality of healthcare and has also suggested a link to health outcomes [2]. Amidst this trend, new health care legislation has recently passed [3] which is aimed at dramatically increasing the number of Americans with access to health care through a government competitor to private insurance. A change of this magnitude can be anticipated to have a significant effect on the provision of health care services and may potentially have many unintended consequences. Many important decision variables come into play including the number of people to insure, the tax levels, and the physician reimbursement levels, to name a few. From basic economic principles it follows that as a profession's profit decreases while its labor workload increases, that profession becomes less and less desirable resulting in a decreased labor force. In this paper we examine potential effects for a number of possible health care policy decisions on physician availability. We developed a system dynamics model to describe the introduction of a government competitor to private insurance. Using this model we begin with a control scenario and proceed to explore six different scenarios obtained by modifying different input parameters. Each scenario reflects a different policy dimension. We report the results and provide a discussion of the dynamics observed.

2. METHODOLOGY AND SCOPE

We developed a system dynamics (SD) model to explore the effects of introducing a government competitor to private insurance. This model aims to describe the inter-relationships between public and private insurance levels to total health care demand, tax levels imposed on physicians, physician profitability, physician workload, overall desirability of the profession, and overall physician supply. SD [5] is a continuous modeling and simulation paradigm based on fluid flows between reservoirs. In contrast to more common simulation paradigms (such as agent-based or discrete-event) which model behavior in terms of the individual entities, SD models behavior at the aggregate level. This makes it a very suitable paradigm for exploring high-level policy decisions. SD has been employed on a wide variety of policy studies in diverse areas such as health care [6], [7], supply chain management [8], and organizational management [9].

In this study we develop a hypothetical control scenario based very loosely on the legislated policy change. In particular, we assume that at the scenario start (Jan. 1 2010), 60% of the US population is covered under private insurance. We assume that starting Jan. 1 2020 the government will mandate that 95% of the population be covered, either via private insurance or via a newly introduced public option. We then test six different treatments in isolation against the control: 1) High Physician Tax, 2) Low Physician Tax, 3) Low Government Reimbursement, 4) High Private Turnover, 5) High Government Dissatisfaction, and 6) High Long-Term Private Insurance Affordability. These treatments correspond to different policy decisions or effects of particular policy decisions, and are obtained by modifying certain input parameters from the control levels. Only treatments in isolation are considered; treatment interactions are not considered. The simulation horizon in each scenario is over a 50-year period, beginning on Jan. 1 2010 and ending on Jan. 1 2060.

This study does not attempt to predict what will happen as a result of the policy changes. Rather, it aims to understand the dynamics that come into play by exploring a hypothetical set of scenarios loosely based on legislated policy changes. Furthermore, system dynamics by nature characterizes system behavior patterns at a very high level of aggregation. It is thus accurate, but not necessarily precise. As a consequence, the results of this study are not to be understood in terms of concrete quantities, but rather in terms of how the system will behave over time relative to the initial conditions.

3. SYSTEM DYNAMICS PRIMER

SD [5] is a continuous modeling and simulation paradigm based on fluid flows between reservoirs (referred to as stocks). Stocks are filled and emptied by inflows and outflows. Differential equations describe flow rates in and out of the stocks contained within the system. Two types of diagrams are commonly used in SD: 1) a causal loop diagram, and 2) a stock-and-flow diagram. The causal loop is purely a conceptual aid that identifies key system components and captures the positive or negative influences that different system components exert on each other. The stock-and-flow diagram is an executable simulation model that identifies the system stocks and flows, as well as specifies the different rates of flow. An example of each type of model is shown in Figure 1 and 2 below:

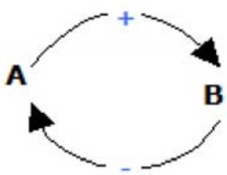


Fig. 1: A causal loop diagram

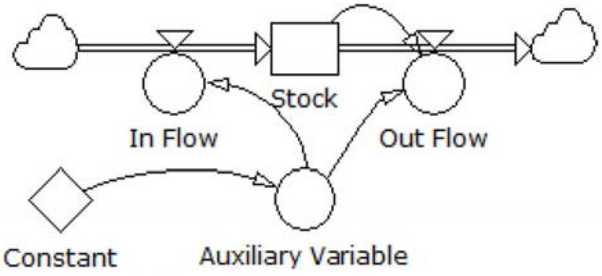


Fig. 2: A stock-and-flow model

In Figure 1, the positive arrow going from A to B indicates that as A increases, B increases. Similarly, the negative arrow indicates that as B increases, A decreases. In Figure 2, stocks are symbolized by boxes and flows by the pipes going in and out of stocks. The clouds represent infinite sources (going in) and sinks (going out). The in- and out-flow rates are determined solely by the flow valves (represented as circles with triangles on top). However, the flow rates can be based on virtually any pertinent component, including stocks, auxiliary variables (represented as circles), and constants (represented by diamonds). This enables fully dynamic relations between system components to be modeled. Arrows represent input into flow rates or auxiliary variables, which are specified as rate equations based on the inputs. For example, 'Auxiliary Variable = Constant * 5/Year', or 'Out Flow = 0.5 * Stock * Auxiliary Variable/Year'.

4. SYSTEM MODEL

We begin the modeling with a causal loop diagram and proceed to use this diagram to develop a full stock-and-flow model. In reality there are a vast number of factors that may have an effect on the healthcare system. However, out of practical concern we restrict the scope of our modeling to capture the interaction of only the key variables we are interested in. The causal loop diagram is shown in Figure 3. In this model the government plan is assumed to be a competitor to private insurance; consequently, they negatively influence each other. However, an increase in either category will increase the demand for healthcare services. An increase in the government insured level will also presumably increase the level of taxes on physicians, which will decrease profits. An increase in the government insured level may also directly decrease profits if the government limits physician reimbursement levels, as has been done with Medicaid [4].

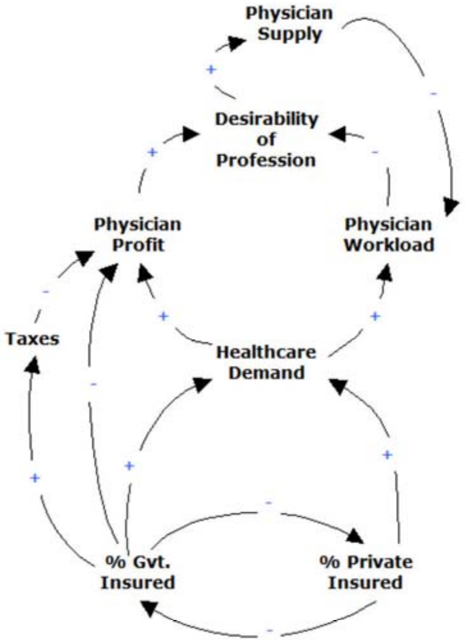


Fig. 3: Causal loop diagram

As healthcare demand increases physician profits will increase through increased revenues; however, their workload will also increase. Intuitively, an increase in profits will positively influence the desirability of the medical profession, while an increased workload will negatively influence it. Similarly, the desirability of the profession will positively influence the physician supply. Finally, an increase to physician supply can be expected to decrease physician workload through load balancing.

The stock-and-flow model is shown in Figure 4. As can be observed, this model has a fairly large number of variables and interactions. It is thus infeasible in this length-limited paper to provide a detailed description of all components. Therefore, we provide a model overview emphasizing aspects most directly related to physician supply. More detail (including equations used for each variable) can be obtained by contacting the authors.

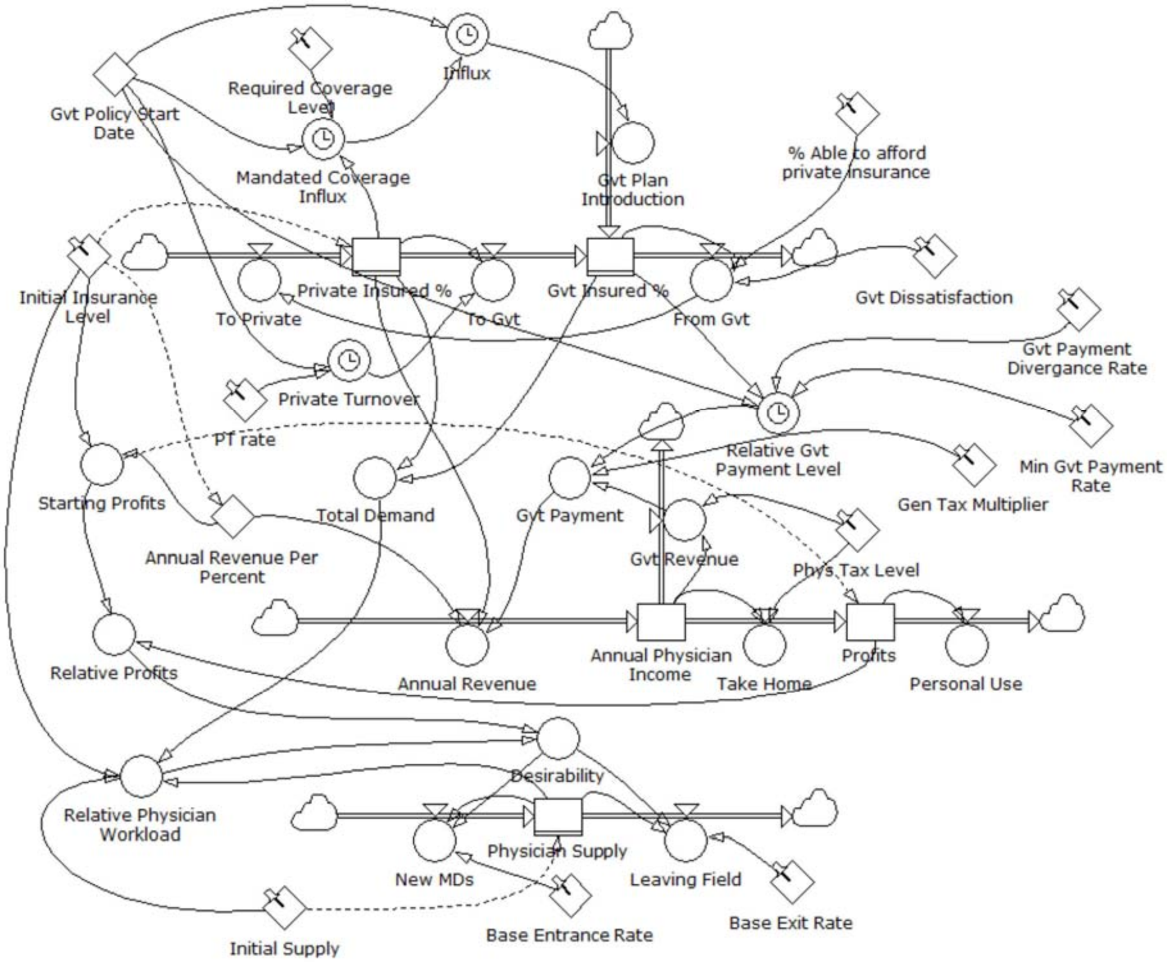


Fig. 4: Stock-and-flow model

In this model, the diamonds with a thumb tack on them are user-adjustable input parameters. The percentages of private and government-insured Americans are modeled as stocks. A certain percentage of the population flows between the private and government insured; however, this does not begin to occur until the policy enactment date (Jan. 1 2020 in this simulation). At this time there will be an influx of newly insured (equal to the difference in mandated insured percentage minus the current private insured level), presumably into the government plan. The auxiliary variables shown with clocks designate that they utilize the start date as input.

Physician incomes and profits are also represented as stocks. In this study we are not interested in the actual amounts of revenues or profits; we are only interested in the changes in these quantities relative to the starting level. We thus use a normalized value for physician income and profit, initially set to 1. As can be seen, physician revenues have two basic sources: public and private insurance. In order to model the potentially limited level of government reimbursement (compared to private insurance) we utilize a "Relative Government Payment Level" variable. This starts the government reimbursement level at 100% of the private level at the start of the policy enactment, and gradually decreases it by a specified percent each year (specified by the "Gvt. Payment Divergence Rate" constant) until the terminal level is reached (specified by the "Min Gvt. Payment Rate" constant). In this model we are only concerned with the effects of taxes on physicians (as opposed to the general population). Thus, we use two constants, "General Tax Multiplier" and "Annual Revenue per Percent" to initialize the normalized revenue to 1. Ultimately, the physician tax level is an independent variable that can be set directly by policy makers. Consequently, we model this as an input constant rather than something determined by emergent system behavior.

Finally, the physician supply is determined by the profession desirability. It is difficult to objectively quantify desirability. However, increased desirability must increase the number of incoming and decrease the number of outgoing physicians. Thus, in terms of this study desirability is defined as the relative profit divided by the relative workload. The relative profit is defined as the current profit level divided by the starting profit level, and the relative workload is defined as the current workload divided by the starting workload. The notion of desirability only makes sense when compared to a starting frame of reference. As a consequence, starting profit and workload levels are set to 1, which initializes the desirability to 1. Values higher than 1 thus indicate a higher-than-starting desirability; values less than 1 indicate a lower-than-starting desirability. Physician supply is represented as a stock, initially set to 300,000. Physician supply inflow and outflow is determined by desirability, a constant base entrance/exit rate, and current physician supply level, where

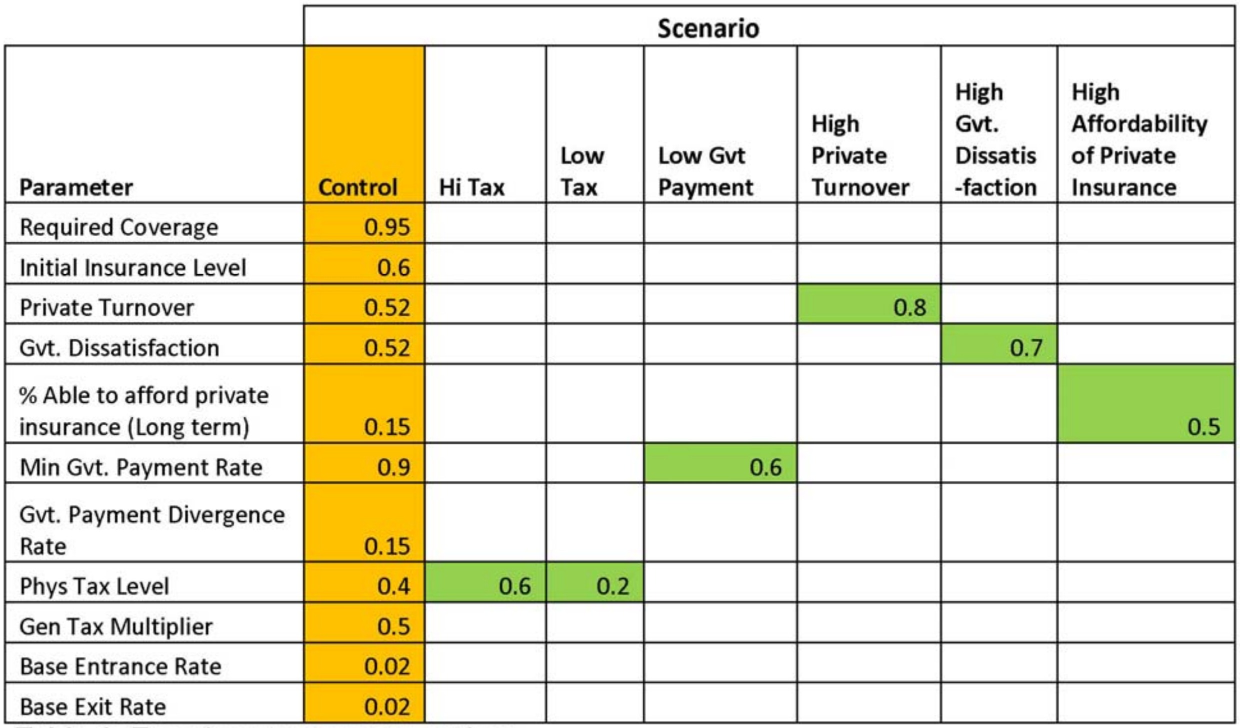
(1) New MDs = Physician Supply * Desirability * Base Entrance Rate

(2) Leaving Field = Physician Supply * (1 / Desirability) * Base Exit Rate

Base entrance/exit rates were set to 2% per year. Thus, an increase in desirability will increase the rate at which new physicians enter the field and slow the exit rate, while a decrease will have exactly the opposite effect.

5. EXPERIMENT DESIGN

We used the Powersim software package to build our simulation model and execute a sequence of scenarios. Using our model, we first developed a relatively stable control scenario. We then proceeded to test six different treatments individually and compare them against the control: 1) High Physician Tax, 2) Low Physician Tax, 3) Low Government Reimbursement, 4) High Private Turnover, 5) High Government Dissatisfaction, and 6) High Long-Time Private Insurance Affordability. These treatments were obtained by modifying certain input parameters from the control levels. We only considered treatments in isolation rather than considering treatment interactions. The parameters for each scenario are shown in Table 1. In non-control scenarios, only parameters that differ from control levels are specified.



	Scenario						
Parameter	Control	Hi Tax	Low Tax	Low Gvt Payment	High Private Turnover	High Gvt. Dissatis -faction	High Affordability of Private Insurance
Required Coverage	0.95						
Initial Insurance Level	0.6						
Private Turnover	0.52				0.8		
Gvt. Dissatisfaction	0.52					0.7	
% Able to afford private insurance (Long term)	0.15						0.5
Min Gvt. Payment Rate	0.9			0.6			
Gvt. Payment Divergence Rate	0.15						
Phys Tax Level	0.4	0.6	0.2				
Gen Tax Multiplier	0.5						
Base Entrance Rate	0.02						
Base Exit Rate	0.02						

Table 1: Experiment parameter settings

In each scenario the same values were used for mandated (required) coverage level, initial insurance level, general tax multiplier, government payment divergence rate, and base entrance/exit rates. Additionally, the normalized revenue and profit levels were set to the initial values of 1, and the physician supply stock was initialized to 300,000. The private insured stock was set to the initial coverage level of 0.6 (corresponding to 60% of the population) and the government insured was initialized to 0. Each scenario had a start date of Jan. 1 2010, with a policy enactment date of Jan. 1 2020. Finally, each scenario had a simulation horizon of 50 years, ending on Jan. 1 2060.

6. RESULTS AND DISCUSSION

The simulation results are summarized in Table 2. As can be seen in the results, a common feature of each scenario is a spike in both the number of physicians and the physician workloads at the onset of the new policy enactment. Because in each scenario the government reimbursement begins at 100% of the private rate, there is a corresponding increase in profits at the start of the policy change. This appears to be due to sheer increase in revenues resulting from the increased demand.

As time moves forward into the future, the diverging effects of different treatments can be seen. The control results in a relatively stable increase in physician supply corresponding to the increase in demand. High physician taxes, low government reimbursement, and high private turnover result in increased physician workloads and decreased physician supply. Of these three, a low government reimbursement rate appears to induce the strongest decline in physician availability. By contrast, low physician taxes, high government dissatisfaction, and high private insurance affordability lead to increases in physician supply and decreased workloads, with the last of these producing the strongest effect.

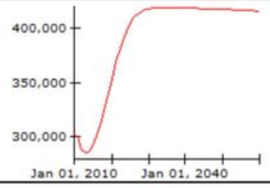
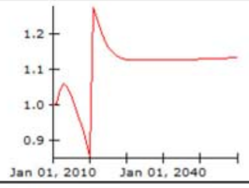
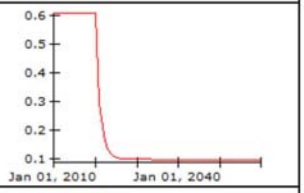
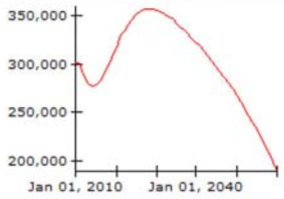
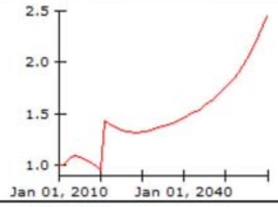
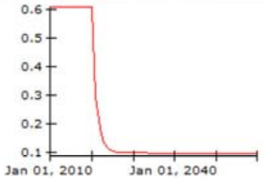
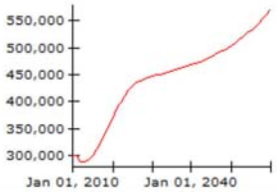
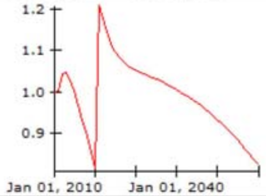
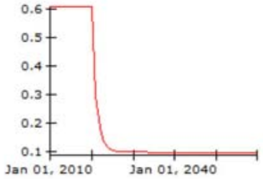
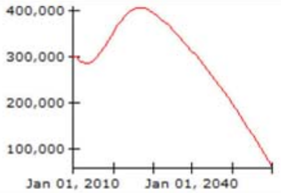
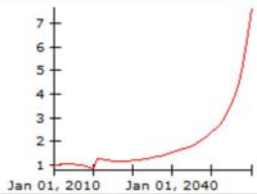
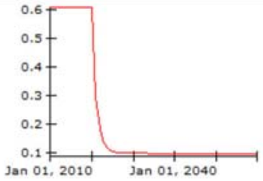
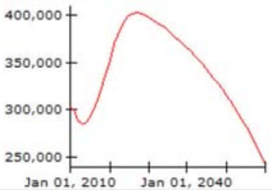
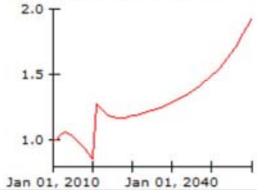
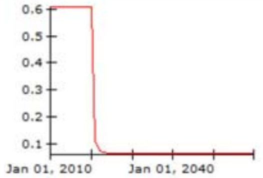
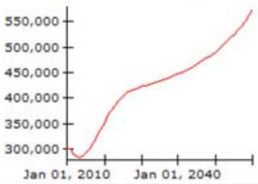
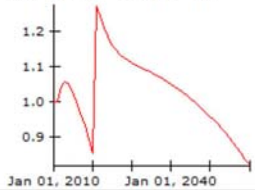
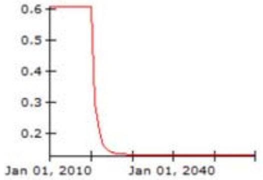
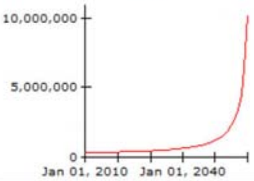
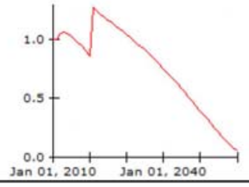
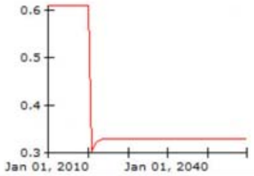
Scenario	Physician Supply	Physician Workload	Private Insured
Control			
High Physician Tax			
Low Physician Tax			
Low Government Reimbursement			
High Private Turnover			
High Government Dissatisfaction			
High Long-Term Private Insurance Affordability			

Table 2: Summary of simulation results

Several of these results appear to be counterintuitive, particularly 1) the high government dissatisfaction leading to increased physician supply, and 2) the disproportionately strong effect of high long-term private insurance affordability. High government dissatisfaction results in a larger number of people able to afford private coverage that choose that option. This increase in privatization leads to higher profits from lower numbers of government-limited reimbursements. The disproportionately strong effect of private insurance affordability is probably due to the control calibration of the "General Tax Multiplier" and "Annual Revenue per Percent" constants, as well as not including other limiting factors (e.g. medical school capacities) in the model. Although the numbers shown are not necessarily precise, they do give a reasonable indication of the emergent system behavior.

In summary, the results indicate that decisions which increase physician profits will increase physician availability, while those that decrease profits will decrease physician availability.

7. CONCLUSIONS

In this paper we have developed a system dynamics model of a government competitor to private insurance. We utilized this model to explore a potential outcomes based on different types of policy decisions or effects. This was accomplished by developing a set of hypothetical scenarios very loosely based on the newly legislated policy changes. We found that in general, decisions that result in increased physician profits will increase physician availability, while those that decrease profits will have the opposite effect.

8. REFERENCES

[1] Marlow, A.E. (1998). "The professional decline of physicians in the era of managed care", *Boston College Dissertations and Theses*. Paper AAI1389161.

[2] Beale, A. and Hernandez, S. (2010) "Patient Reports of the Quality of Care in Community Health Centers: The Importance of Having a Regular Provider", *Journal of Health Care for the Poor and Underserved*, Vol. 21, No. 2, pp. 591-605.

[3] H.R. 4872 (2010). *The Reconciliation Act of 2010.*

[4] H.R. 3962 ENR (2010). *Preservation of Access to Care for Medicare Beneficiaries and Pension Relief Act of 2010.*

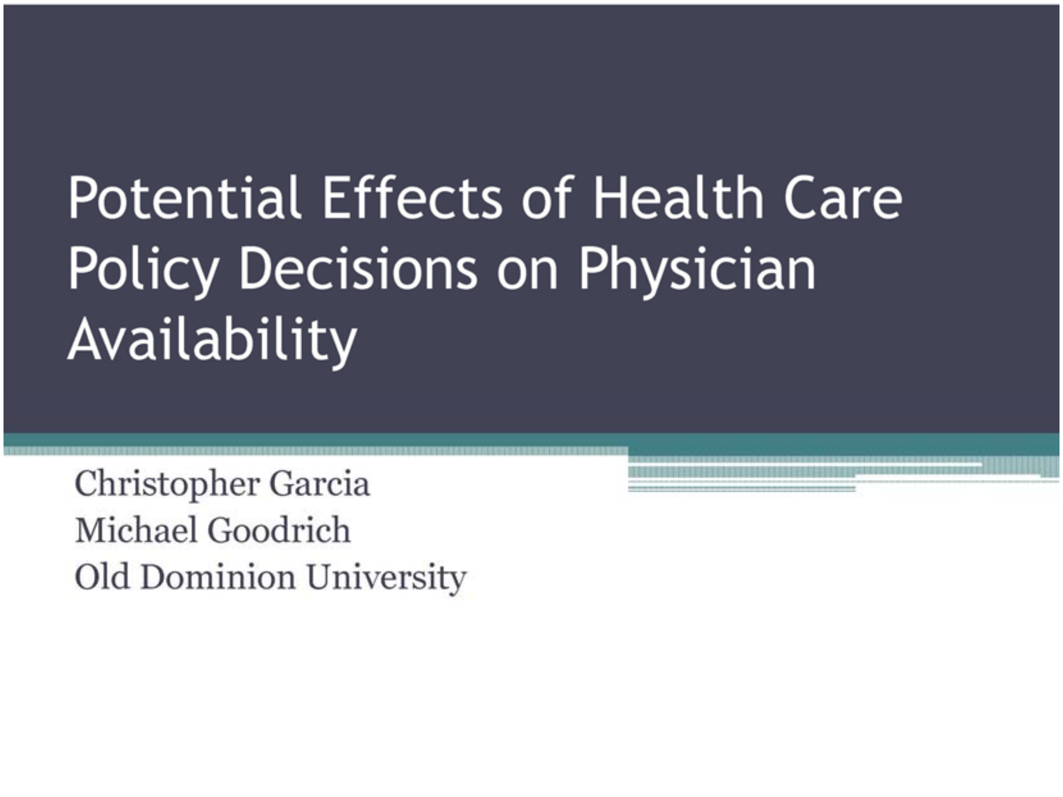
[5] Sternman, J.D. (2000). *Business Dynamics: Systems Thinking and Modeling for a Complex World.* New York: McGraw Hill.

[6] Sardiwal, S. (2007). "Conceptualization and formulation of a UK health and social care system using System Dynamics", Proceedings of the 25th Annual Conference of the System Dynamics Society, Boston, MA, July 29-August 2, 2007.

[7] Koshio, A. and Akiyama, M. (2008). "Physician's burning out and Human resource crisis in Japanese Hospital: Management for sustaining medical services in Japan", Proceedings of the 26th Annual Conference of the System Dynamics Society, Athens, Greece, July 20-24, 2008.

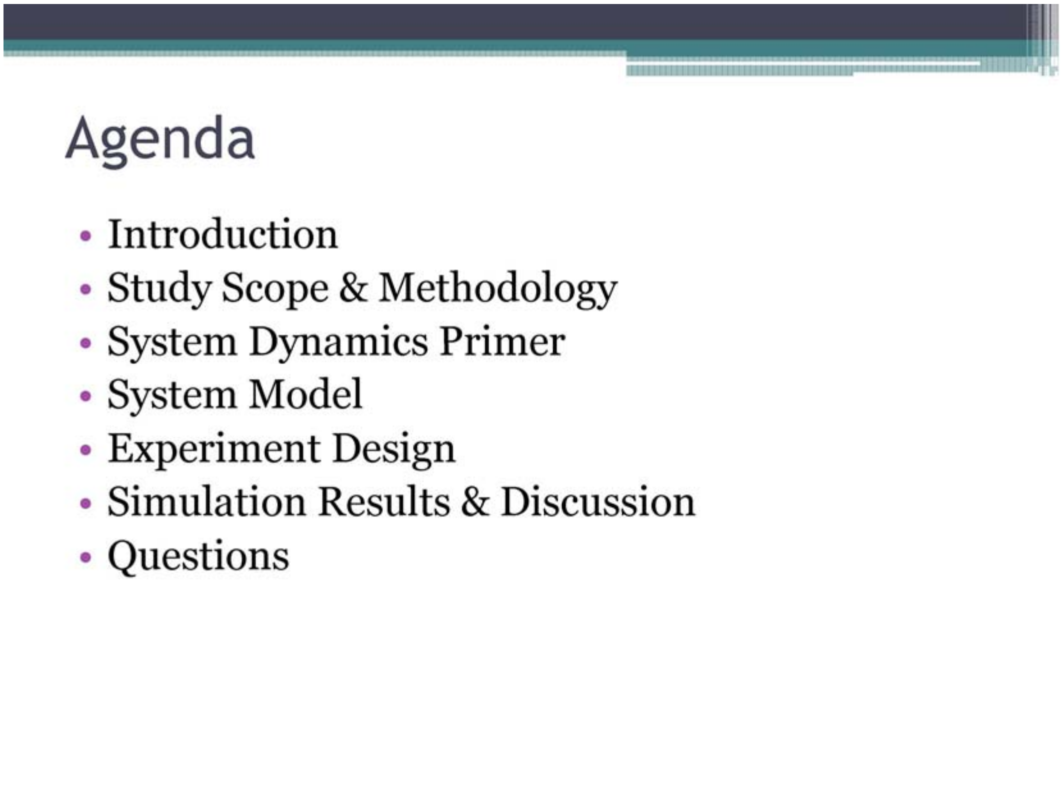
[8] Bijulal, D. and Venkateswaran, J. (2008). "Closed-Loop Supply Chain Stability under Different Production-Inventory Policies", Proceedings of the 26th Annual Conference of the System Dynamics Society, Athens, Greece, July 20-24, 2008.

[9] Heiko, B. (2008). "Legitimacy Crises and Organizational Behavior", Proceedings of the 26th Annual Conference of the System Dynamics Society, Athens, Greece, July 20-24, 2008.



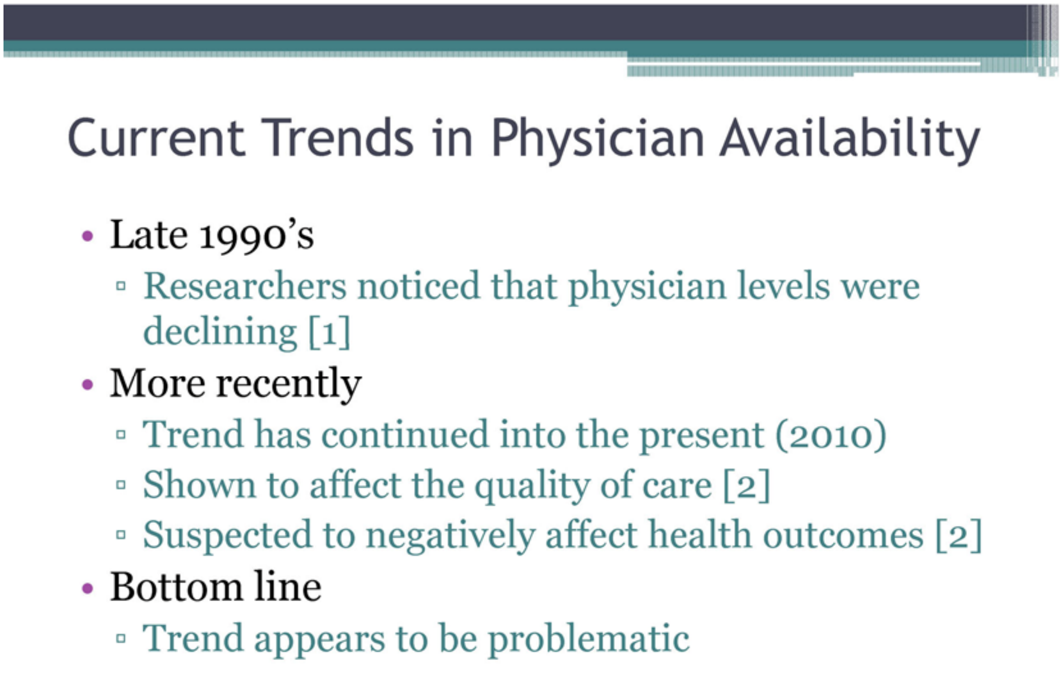
Potential Effects of Health Care Policy Decisions on Physician Availability

Christopher Garcia
Michael Goodrich
Old Dominion University



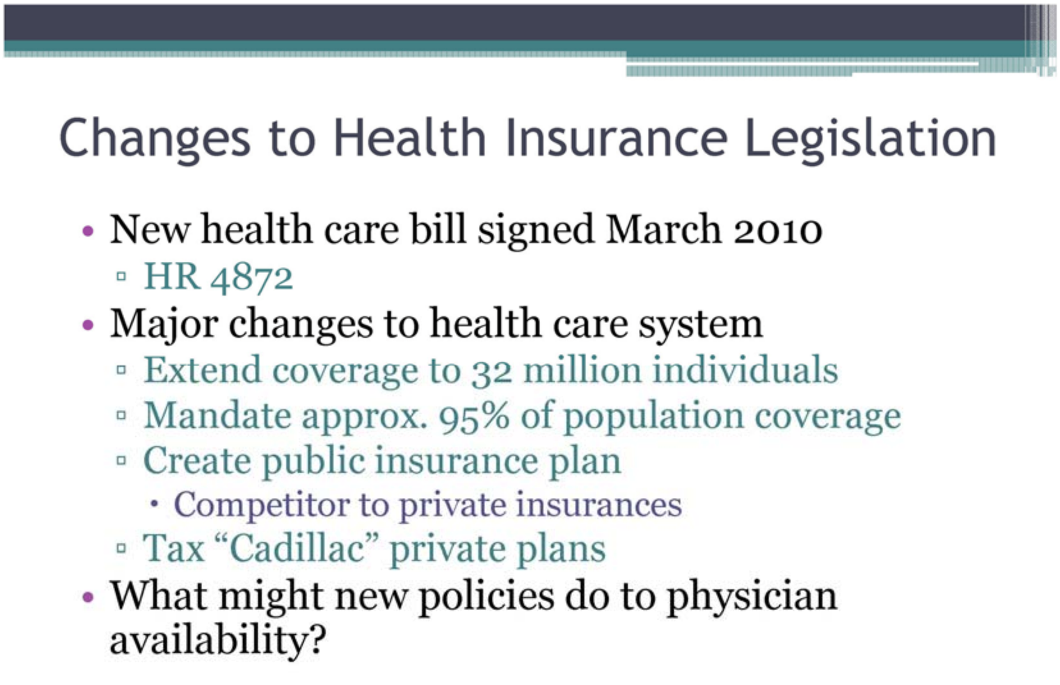
Agenda

- Introduction
- Study Scope & Methodology
- System Dynamics Primer
- System Model
- Experiment Design
- Simulation Results & Discussion
- Questions



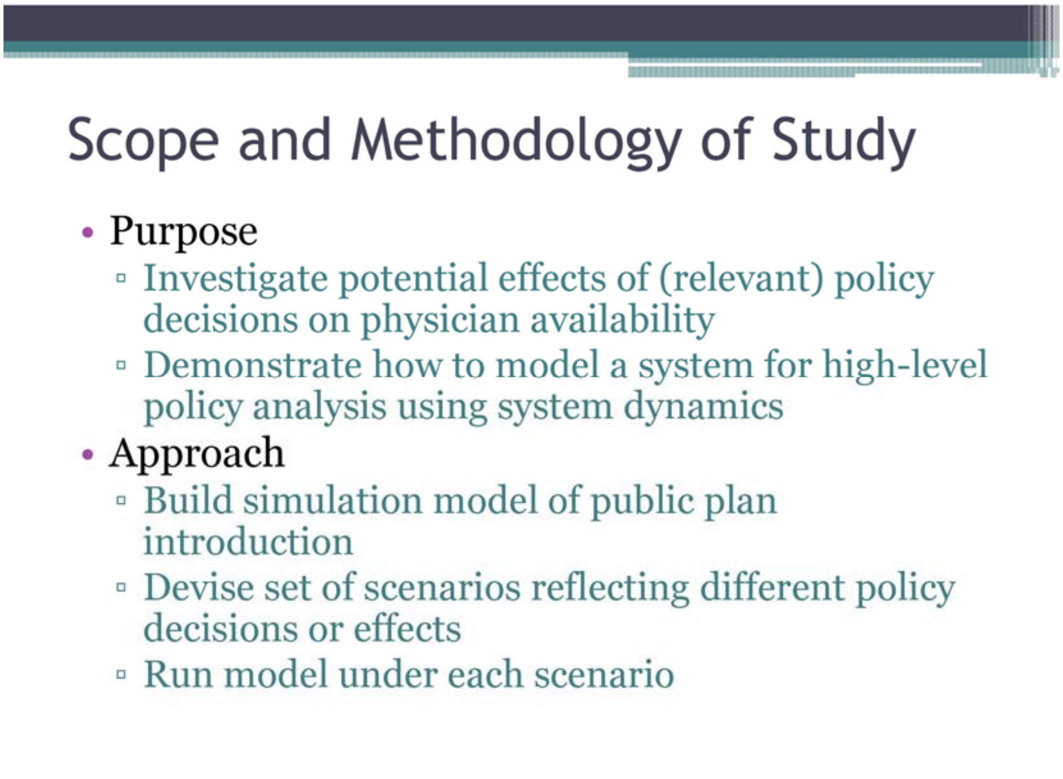
Current Trends in Physician Availability

- Late 1990's
  - Researchers noticed that physician levels were declining [1]
- More recently
  - Trend has continued into the present (2010)
  - Shown to affect the quality of care [2]
  - Suspected to negatively affect health outcomes [2]
- Bottom line
  - Trend appears to be problematic



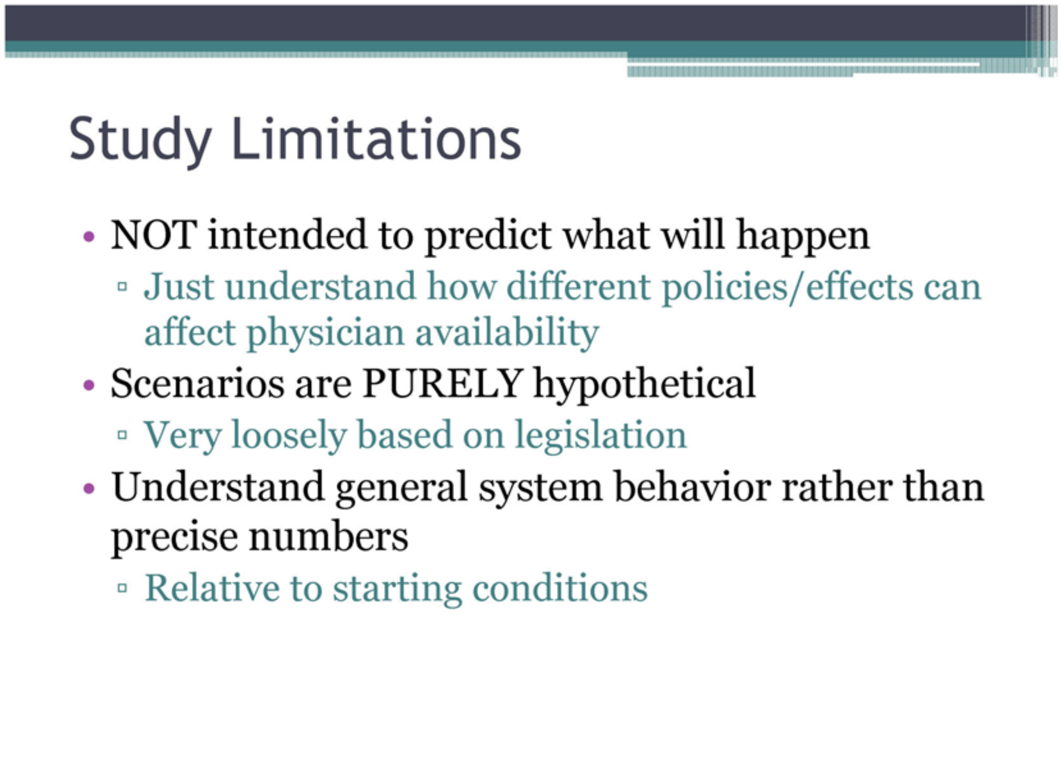
Changes to Health Insurance Legislation

- New health care bill signed March 2010
  - HR 4872
- Major changes to health care system
  - Extend coverage to 32 million individuals
  - Mandate approx. 95% of population coverage
  - Create public insurance plan
    - Competitor to private insurances
  - Tax "Cadillac" private plans
- What might new policies do to physician availability?



Scope and Methodology of Study

- Purpose
  - Investigate potential effects of (relevant) policy decisions on physician availability
  - Demonstrate how to model a system for high-level policy analysis using system dynamics
- Approach
  - Build simulation model of public plan introduction
  - Devise set of scenarios reflecting different policy decisions or effects
  - Run model under each scenario

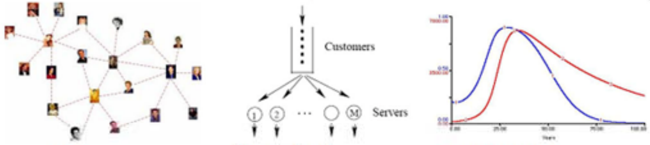


Study Limitations

- NOT intended to predict what will happen
  - Just understand how different policies/effects can affect physician availability
- Scenarios are PURELY hypothetical
  - Very loosely based on legislation
- Understand general system behavior rather than precise numbers
  - Relative to starting conditions

System Dynamics Primer

- SD is a continuous, macro-level M&S paradigm



Agent-Based
Highly Discrete
Micro, Agent-Level

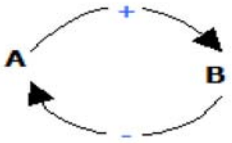
Discrete-Event
Discrete
Mid, Interaction Level

System Dynamics
Continuous
Macro, Aggregate Level

- Model a system of differential equations
- Basic Constructs:
  - Stocks (Reservoirs)
  - Flows (Pipes going in and out of stocks)

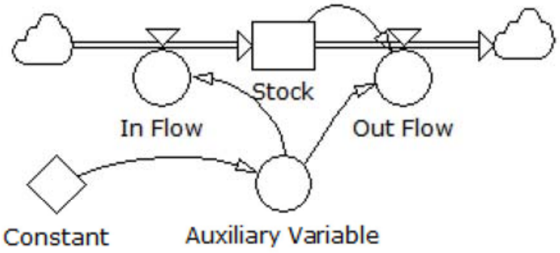
Causal Loop Diagram

- Preliminary, purely a conceptual aid
  - Identifies important system components
  - Identifies positive & negative mutual influences



- Increase in A ➔ Increase in B
- Increase in B ➔ *Decrease* in A

Stock-and-Flow Model

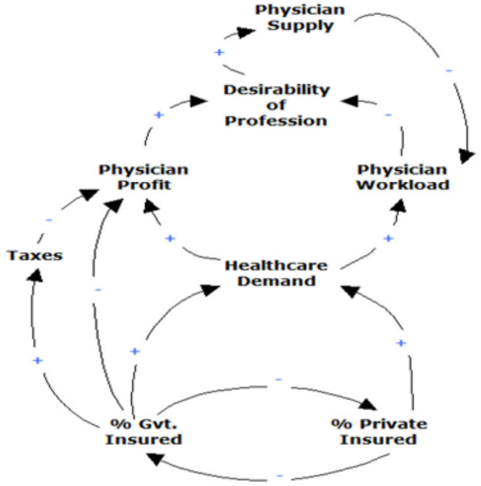


- Executable simulation model
- Simulate a system as “Fluid Flows” over time
- Model flows & auxiliaries with equations
  - Auxiliary Variable = Constant * 5
  - Out Flow = 0.5 * Stock * Auxiliary Variable/Year
  - Etc.

Health Care System Modeling

- Basic Considerations
  - Concerned ONLY with effect of policy decisions on physician availability
  - Identify MAIN factors at play
  - Many things affect outcome, but DO NOT clutter model with non-key pieces
- Basic Assumptions
  - Would-be doctors have many career possibilities
  - Insufficiently rewarding profession will cause them to go elsewhere

Health Care Model: Causal Loop



Physician Supply

Desirability of Profession

Physician Profit

Physician Workload

Taxes

Healthcare Demand

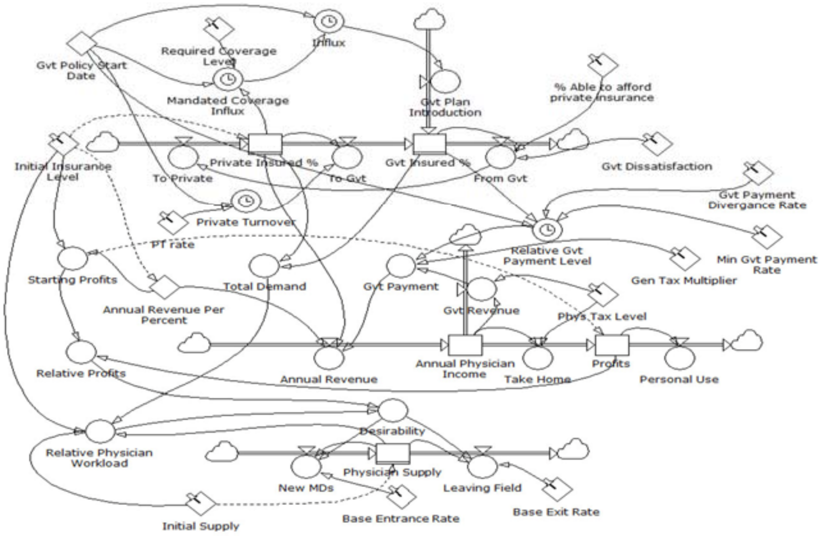
% Gvt. Insured

% Private Insured

HC Stock-and-Flow: Main Aspects

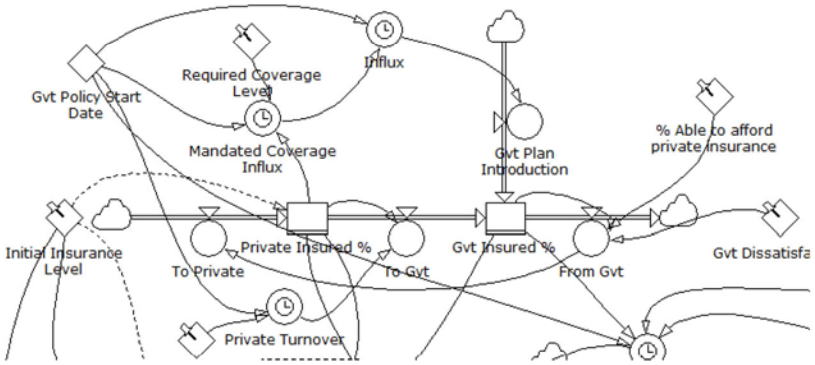
- Three basic “subsystems”:
  - Private/Government Insurance Levels
  - Revenue & Profit Flow
  - Physician Supply & Desirability
- All subsystems influence each other
- Fairly complex model

Full HC Stock-and-Flow Model



Insurance Level Subsystem

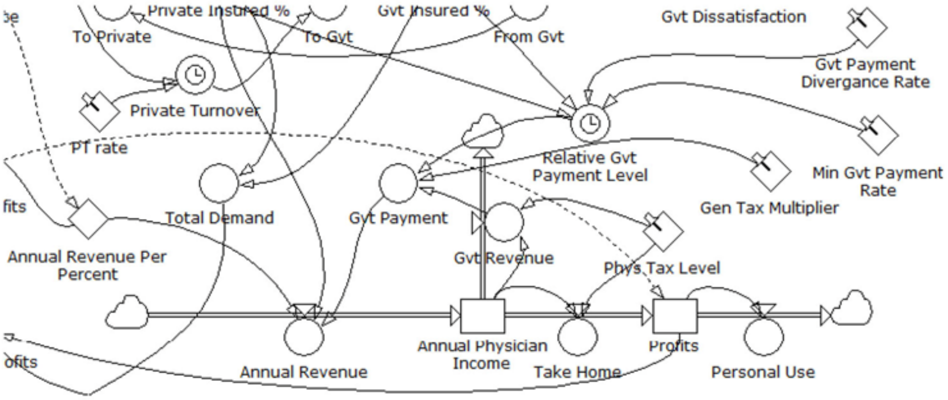
- Mandated 95% of population flows back back & forth starting at policy enactment date



Revenue Subsystem

- Physician income has 2 sources: public, private
- Consider only effect of *physician* taxes in this model
  - Controllable input, not emergent behavior
- Phys. tax level affects Profits & Amt. that goes back to government
  - Use "General Tax Multiplier" to represent rest of population, get sustainable tax revenues
- Government may limit reimbursements
  - Medicare-style limitations
  - Assume that Government starts out at 100% reimbursement at policy start
  - Gradually decrease until terminal level reached

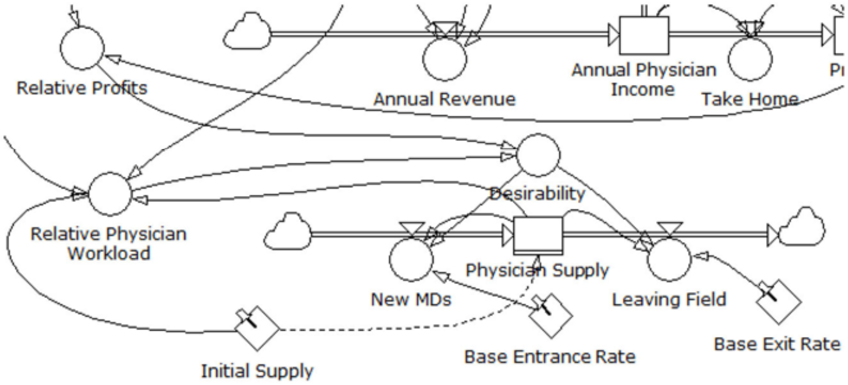
Revenue Subsystem (cont.)

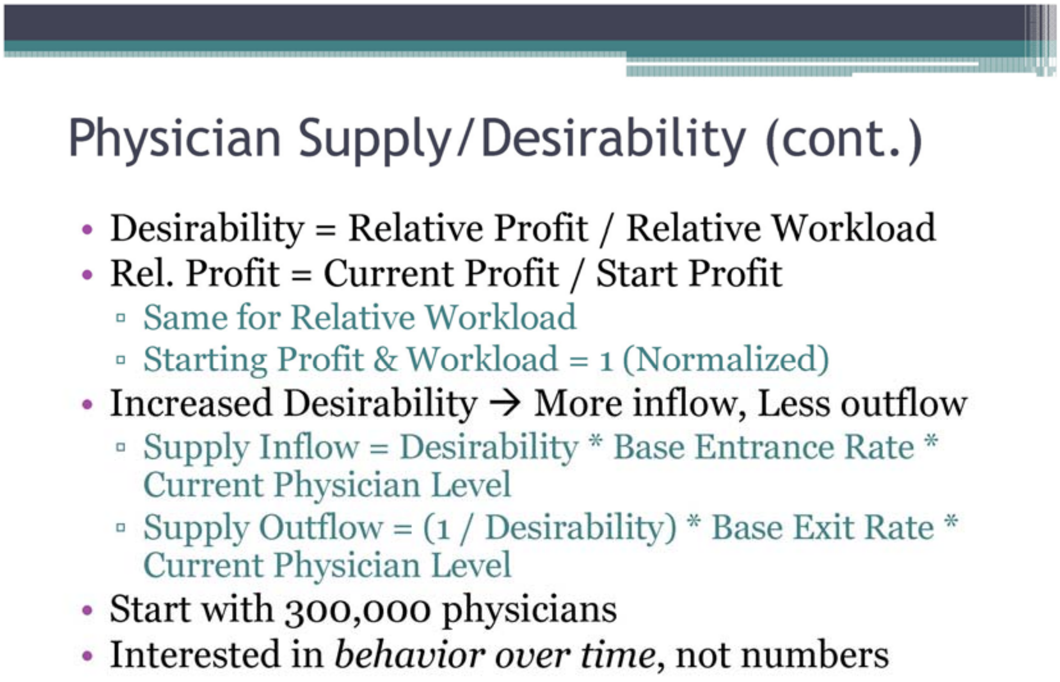


Revenue Subsystem (cont.)

- Interested only in *relative* change in income/profits over time
    - Do levels go up or down over time?
    - Not interested in actual numbers
- Set initial revenue & profits to normalized value of 1
    - > 1 ➔ Higher than start
    - < 1 ➔ Lower than start

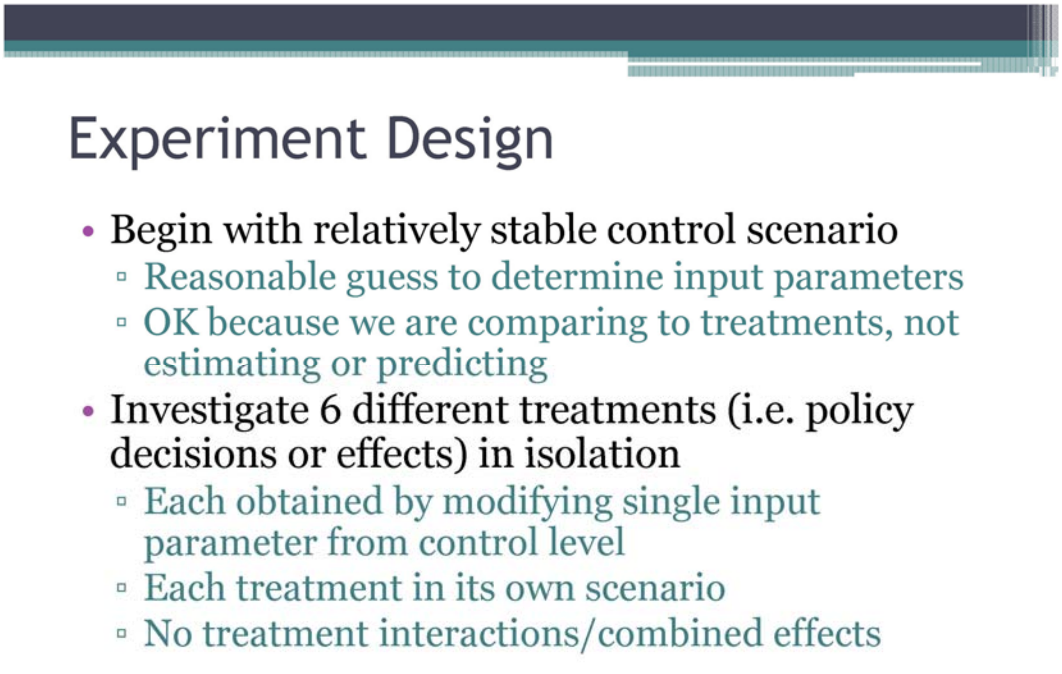
Physician Supply/Desirability





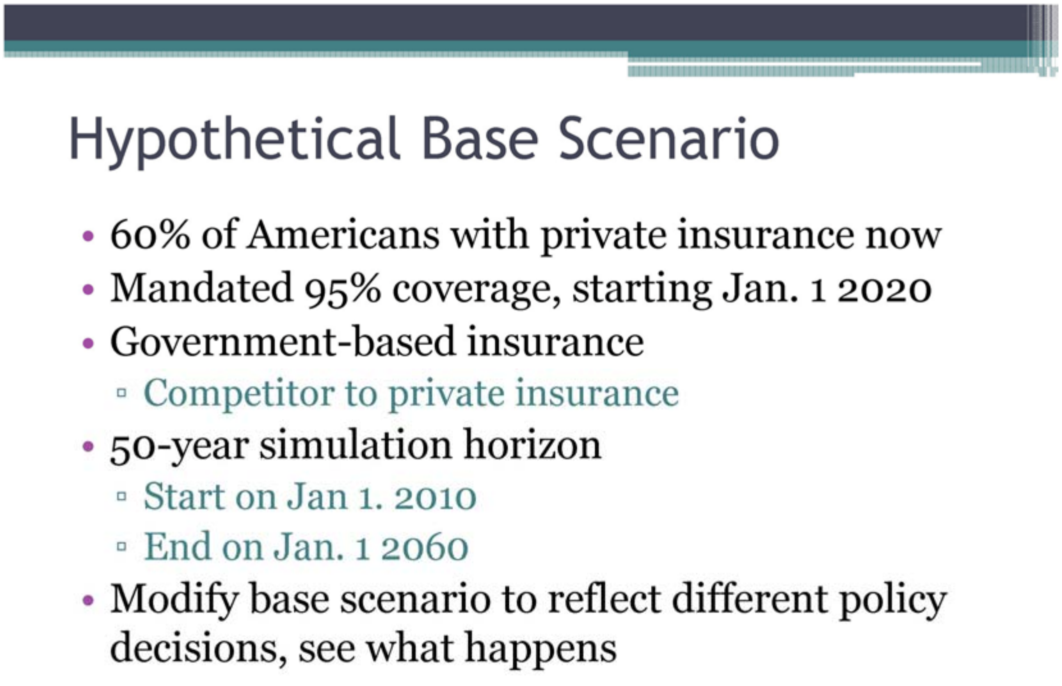
Physician Supply/Desirability (cont.)

- Desirability = Relative Profit / Relative Workload
- Rel. Profit = Current Profit / Start Profit
  - Same for Relative Workload
  - Starting Profit & Workload = 1 (Normalized)
- Increased Desirability → More inflow, Less outflow
  - Supply Inflow = Desirability * Base Entrance Rate * Current Physician Level
  - Supply Outflow = (1 / Desirability) * Base Exit Rate * Current Physician Level
- Start with 300,000 physicians
- Interested in *behavior over time*, not numbers



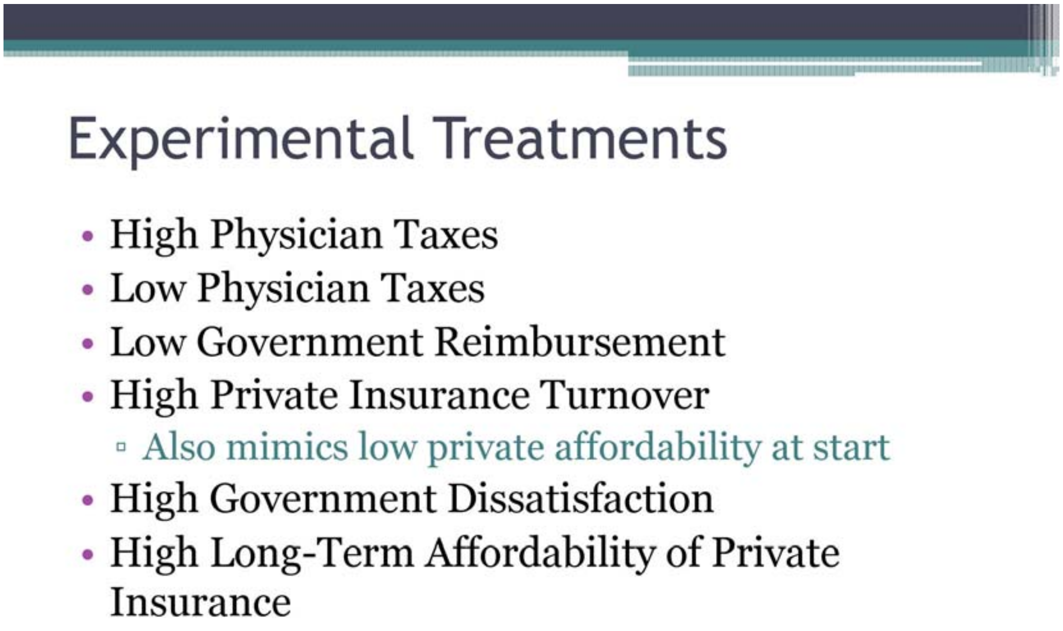
Experiment Design

- Begin with relatively stable control scenario
  - Reasonable guess to determine input parameters
  - OK because we are comparing to treatments, not estimating or predicting
- Investigate 6 different treatments (i.e. policy decisions or effects) in isolation
  - Each obtained by modifying single input parameter from control level
  - Each treatment in its own scenario
  - No treatment interactions/combined effects



Hypothetical Base Scenario

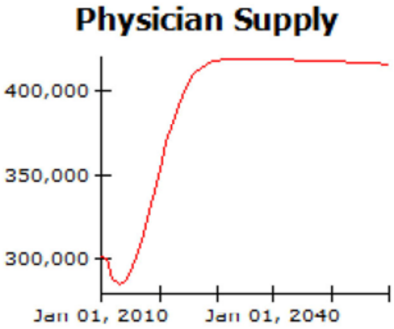
- 60% of Americans with private insurance now
- Mandated 95% coverage, starting Jan. 1 2020
- Government-based insurance
  - Competitor to private insurance
- 50-year simulation horizon
  - Start on Jan 1. 2010
  - End on Jan. 1 2060
- Modify base scenario to reflect different policy decisions, see what happens

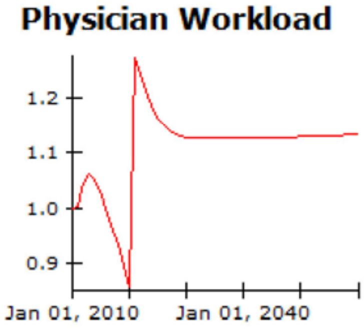


Experimental Treatments

- High Physician Taxes
- Low Physician Taxes
- Low Government Reimbursement
- High Private Insurance Turnover
  - Also mimics low private affordability at start
- High Government Dissatisfaction
- High Long-Term Affordability of Private Insurance

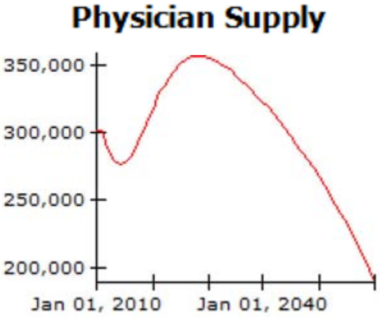
Control Scenario

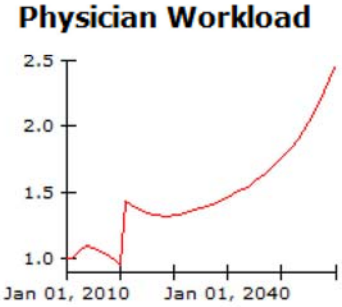

Physician Supply

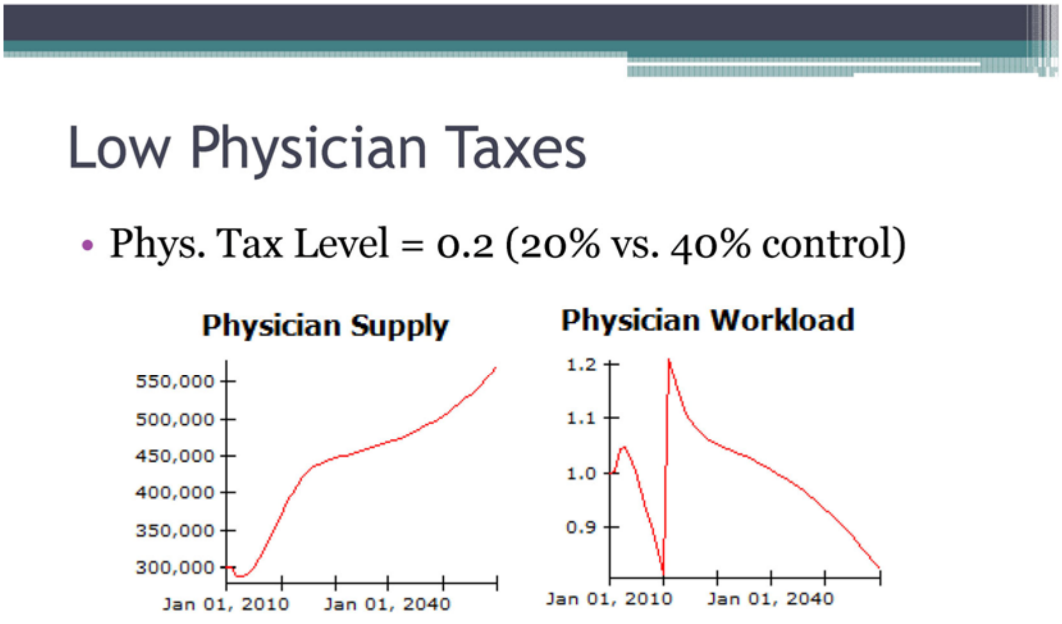

Physician Workload

High Physician Taxes

- Phys. Tax Level = 0.6 (60% vs. 40% control)


Physician Supply


Physician Workload

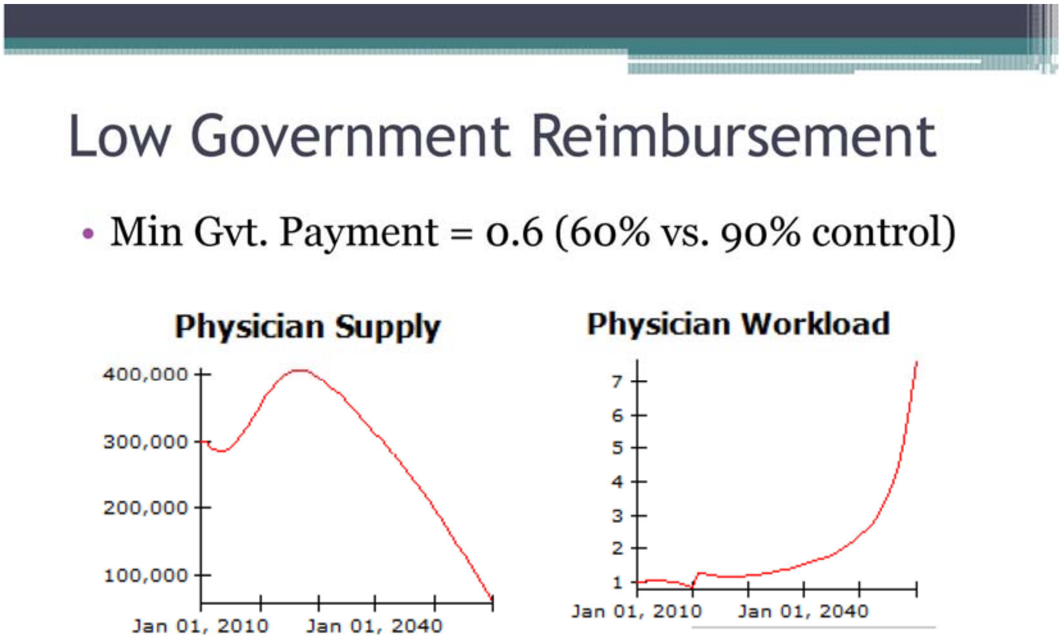


Low Physician Taxes

- Phys. Tax Level = 0.2 (20% vs. 40% control)

Physician Supply

Physician Workload

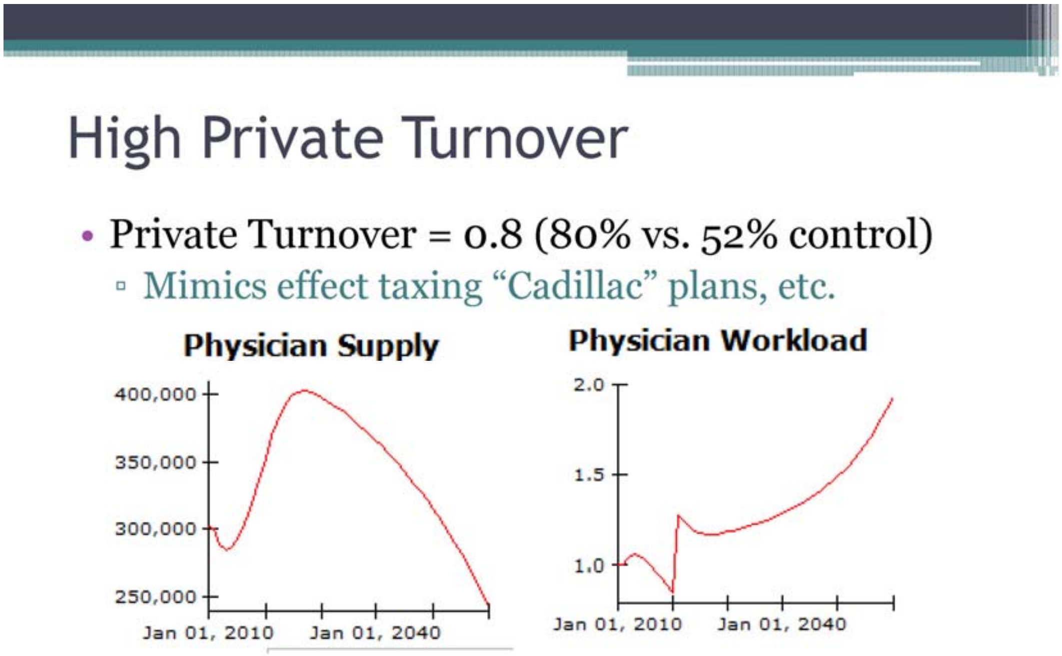


Low Government Reimbursement

- Min Gvt. Payment = 0.6 (60% vs. 90% control)

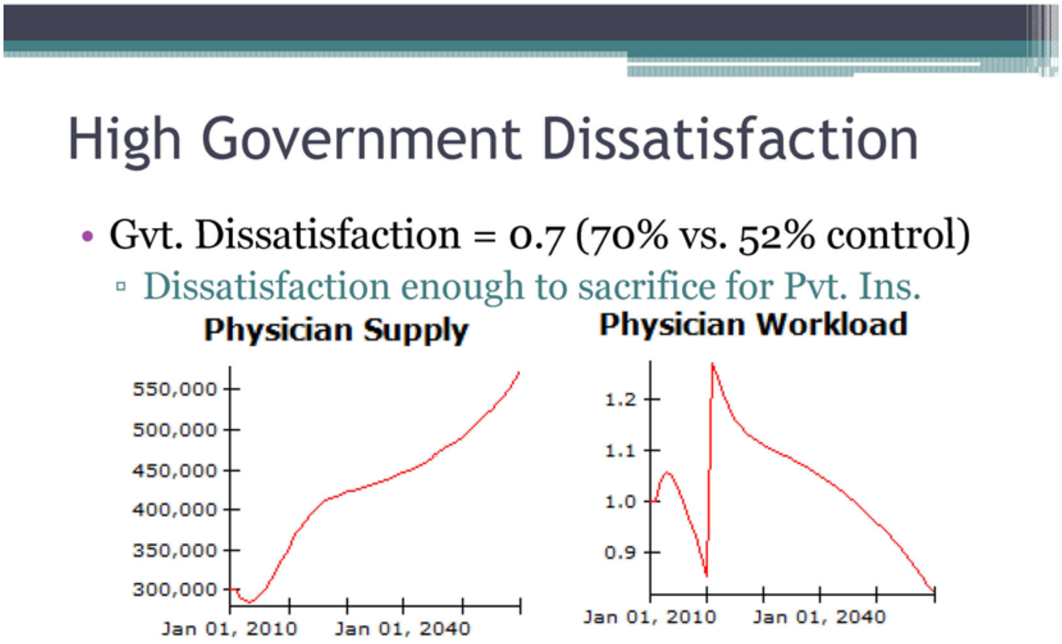
Physician Supply

Physician Workload



High Private Turnover

- Private Turnover = 0.8 (80% vs. 52% control)
  - Mimics effect taxing "Cadillac" plans, etc.



High Government Dissatisfaction

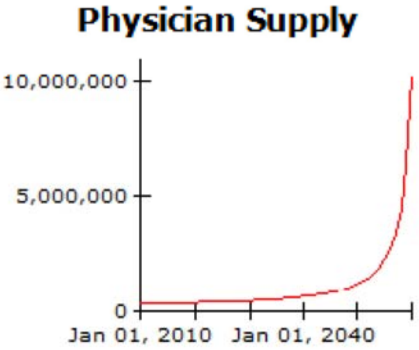
- Gvt. Dissatisfaction = 0.7 (70% vs. 52% control)
  - Dissatisfaction enough to sacrifice for Pvt. Ins.

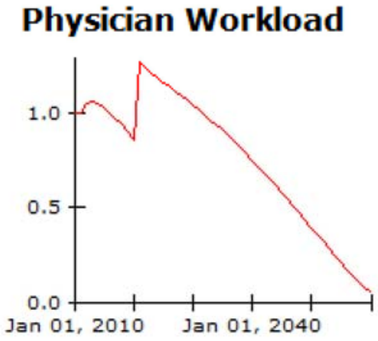
Explanation: High Gvt. Dissatisfaction

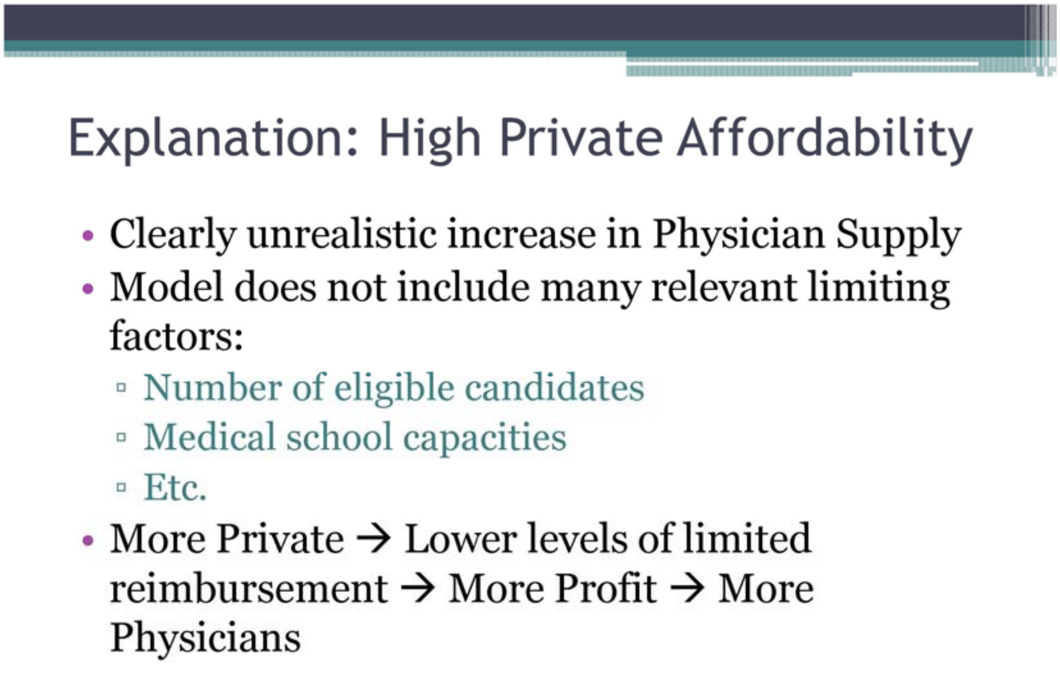
- Counterintuitive result
- Assumes government limits reimbursement to 90% in long run
- More Gvt. Dissatisfaction → More Private
- More Private → Lower levels of limited reimbursement
- Bottom line:
  - Increase in supply linked to limited reimbursement

High Private Insurance Affordability

- Able to Afford Private Insurance = 0.5 (vs. 15%)

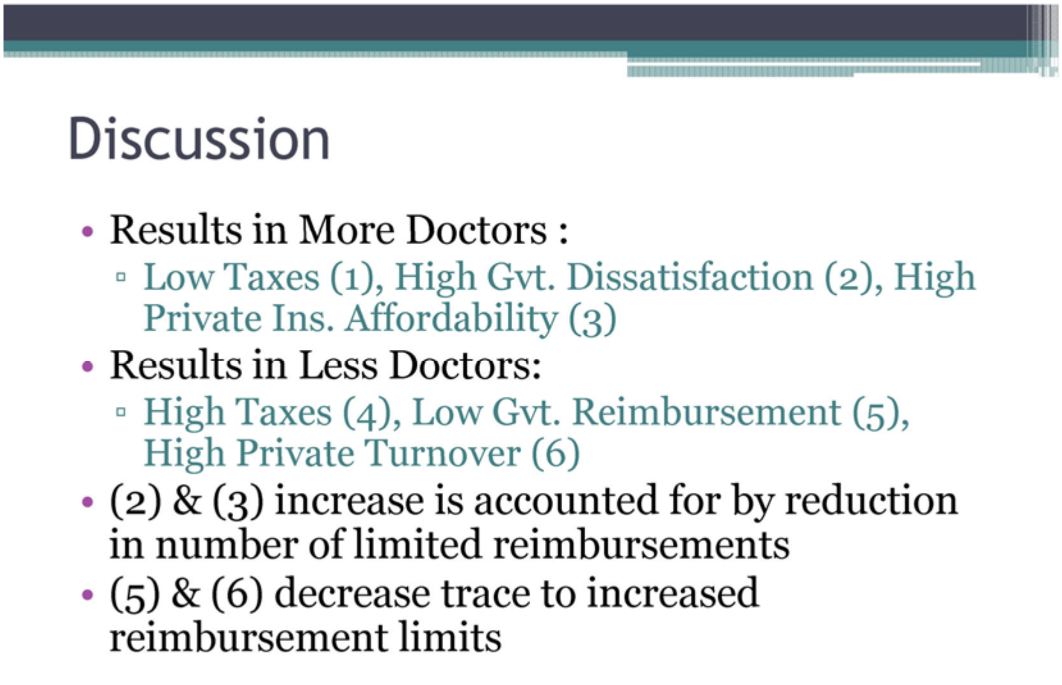






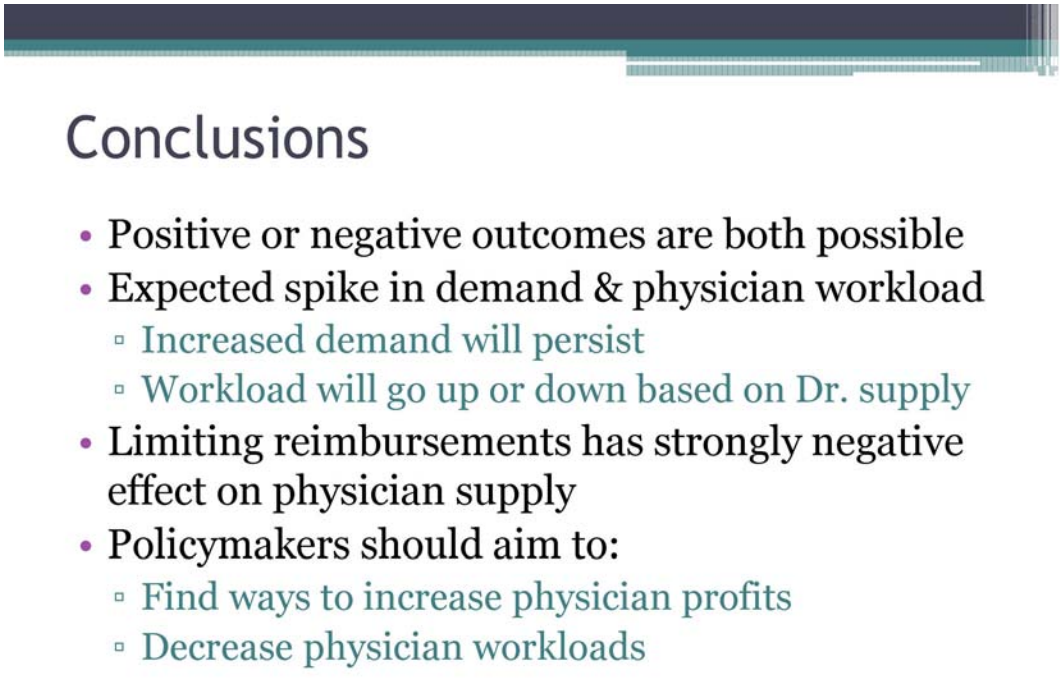
Explanation: High Private Affordability

- Clearly unrealistic increase in Physician Supply
- Model does not include many relevant limiting factors:
  - Number of eligible candidates
  - Medical school capacities
  - Etc.
- More Private → Lower levels of limited reimbursement → More Profit → More Physicians



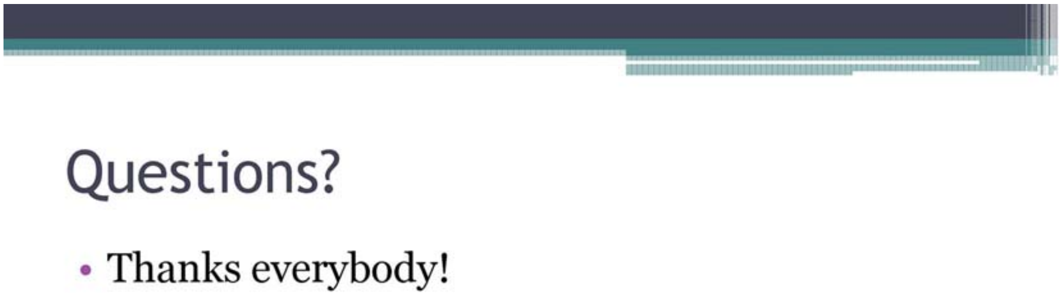
Discussion

- Results in More Doctors :
  - Low Taxes (1), High Gvt. Dissatisfaction (2), High Private Ins. Affordability (3)
- Results in Less Doctors:
  - High Taxes (4), Low Gvt. Reimbursement (5), High Private Turnover (6)
- (2) & (3) increase is accounted for by reduction in number of limited reimbursements
- (5) & (6) decrease trace to increased reimbursement limits



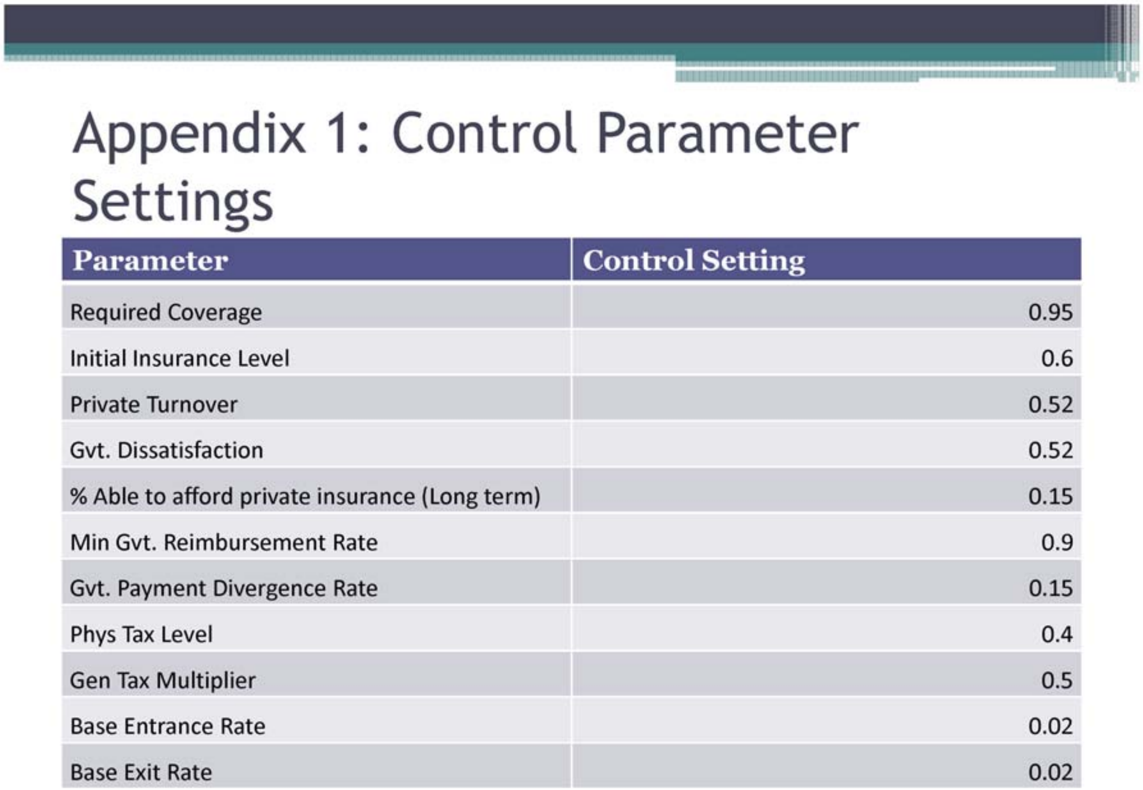
Conclusions

- Positive or negative outcomes are both possible
- Expected spike in demand & physician workload
  - Increased demand will persist
  - Workload will go up or down based on Dr. supply
- Limiting reimbursements has strongly negative effect on physician supply
- Policymakers should aim to:
  - Find ways to increase physician profits
  - Decrease physician workloads



Questions?

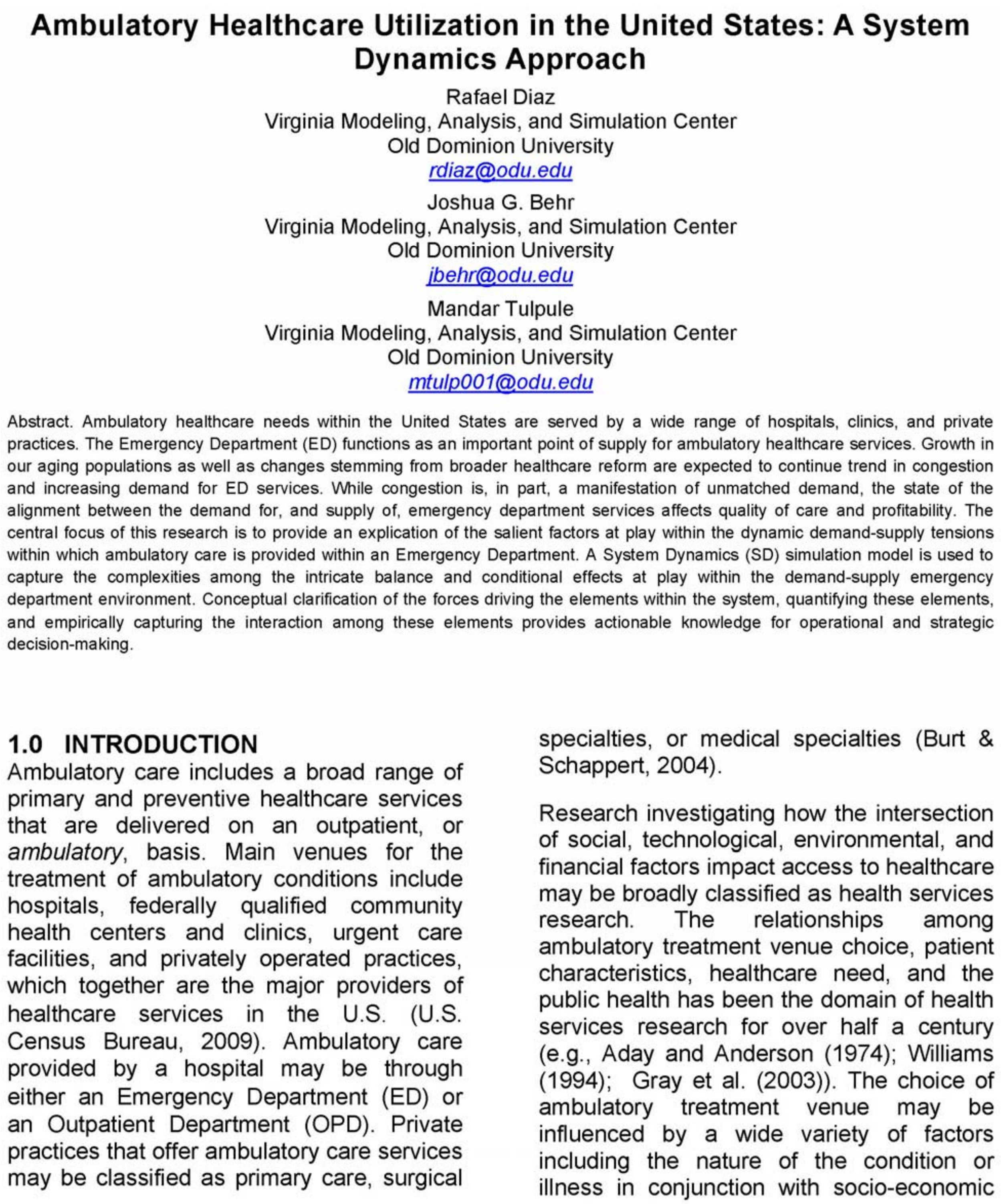
- Thanks everybody!



Appendix 1: Control Parameter Settings

Parameter	Control Setting
Required Coverage	0.95
Initial Insurance Level	0.6
Private Turnover	0.52
Gvt. Dissatisfaction	0.52
% Able to afford private insurance (Long term)	0.15
Min Gvt. Reimbursement Rate	0.9
Gvt. Payment Divergence Rate	0.15
Phys Tax Level	0.4
Gen Tax Multiplier	0.5
Base Entrance Rate	0.02
Base Exit Rate	0.02

3.7 Ambulatory Healthcare Utilization in the United States: A System Dynamics Approach



Ambulatory Healthcare Utilization in the United States: A System Dynamics Approach

Rafael Diaz
Virginia Modeling, Analysis, and Simulation Center
Old Dominion University
rdiaz@odu.edu

Joshua G. Behr
Virginia Modeling, Analysis, and Simulation Center
Old Dominion University
jbehr@odu.edu

Mandar Tulpule
Virginia Modeling, Analysis, and Simulation Center
Old Dominion University
mtulp001@odu.edu

Abstract. Ambulatory healthcare needs within the United States are served by a wide range of hospitals, clinics, and private practices. The Emergency Department (ED) functions as an important point of supply for ambulatory healthcare services. Growth in our aging populations as well as changes stemming from broader healthcare reform are expected to continue trend in congestion and increasing demand for ED services. While congestion is, in part, a manifestation of unmatched demand, the state of the alignment between the demand for, and supply of, emergency department services affects quality of care and profitability. The central focus of this research is to provide an explication of the salient factors at play within the dynamic demand-supply tensions within which ambulatory care is provided within an Emergency Department. A System Dynamics (SD) simulation model is used to capture the complexities among the intricate balance and conditional effects at play within the demand-supply emergency department environment. Conceptual clarification of the forces driving the elements within the system, quantifying these elements, and empirically capturing the interaction among these elements provides actionable knowledge for operational and strategic decision-making.

1.0 INTRODUCTION

Ambulatory care includes a broad range of primary and preventive healthcare services that are delivered on an outpatient, or *ambulatory*, basis. Main venues for the treatment of ambulatory conditions include hospitals, federally qualified community health centers and clinics, urgent care facilities, and privately operated practices, which together are the major providers of healthcare services in the U.S. (U.S. Census Bureau, 2009). Ambulatory care provided by a hospital may be through either an Emergency Department (ED) or an Outpatient Department (OPD). Private practices that offer ambulatory care services may be classified as primary care, surgical specialties, or medical specialties (Burt & Schappert, 2004).

Research investigating how the intersection of social, technological, environmental, and financial factors impact access to healthcare may be broadly classified as health services research. The relationships among ambulatory treatment venue choice, patient characteristics, healthcare need, and the public health has been the domain of health services research for over half a century (e.g., Aday and Anderson (1974); Williams (1994); Gray et al. (2003)). The choice of ambulatory treatment venue may be influenced by a wide variety of factors including the nature of the condition or illness in conjunction with socio-economic and spatial factors, all of which influence the demand for ambulatory care services (Behr, Drivers of Emergency Department Utilization: A Systematic Study of the Individual's Decision Calculus to Select Healthcare Services, 2008).

It is clear that the nature of illness and chronic disease within our society creates the *demand* for treatment. The system of ambulatory care venues constitutes the *supply* of treatment meant to satisfy this demand. The capacity to supply treatment is finite and is a product of the availability of health care professionals and ambulatory care facilities. It is of national interest to match the demand for treatment with the availability and access to ambulatory treatment venues. Ensuring that all population segments, especially those that have traditionally been underserved, uninsured, or underinsured, have access to a reliable ambulatory care system is often the focus of public policy makers.

The complexity of factors, many of which are dynamically inter-related, may frustrate reaching demand-supply equilibrium. For example, demand for ambulatory care may be initially reduced by a condition of non- or under-insurance stemming from socio-economic conditions; potential patients may delay seeking ambulatory care (Behr & Diaz, 2010), yet delayed or non-treatment may be the catalyst for higher-acuity conditions that later may placed increased demand on the system. On the supply side, congestion in Emergency Departments may be a product of staffing and nursing issues and this, in turn, may impact timely treatment (Carr, Kaye, Wiebe, Gracias, Schwab, & Reilly, 2007). The interplay between demand and supply conditions the availability, affordability, congestion and quality of ambulatory care.

We assert that a meaningful understanding of the ambulatory care system may be derived from a holistic and encompassing approach that considers both demand and supply factors. Adequately identifying the salient supply-demand factors and realistically capturing the dynamics among critical factors are essential requirements in the development of a model of the ambulatory care system. Specifying within the model an appropriate combination of both demand and supply variables that captures the real-world system will then allow us to measure changes in variables that approach a sustainable healthcare system, one in which we have a viable balance between demand and supply.

The purpose of this study is to construct a high level, calibrated model for ambulatory healthcare utilization in the U.S. This research offers a System Dynamics (SD) simulation model that captures from an Emergency Department perspective the complexities associated with the interactions among these factors. This model demonstrates that it is possible to capture the supply and demand complexities that define our ambulatory care system.

The validity of the simulation model is demonstrated by comparing actual data from United States Ambulatory healthcare statistics with data produced by the simulation. Specifically, this study models the overall ambulatory healthcare demand and utilization trends in the United States between 1999 and 2006. The model and simulation mimics the current national-level dynamics of the ambulatory system with emphasis on emergency management and provides insights that may inform policy making.

Following this Introduction, Section 2 describes the research question, Section 3 presents an overview of our modeling and simulation approach and Section 4 presents and validates the results. Finally, the last Section discusses conclusion and the potential for the model to be extended to other applications.

2.0 USING SYSTEM DYNAMICS TO SIMULATE EMERGENCY DEPARTMENTS

The critical role of Emergency Departments in the delivery of healthcare in the United States makes them an important subject for research and analysis. Simulation may be employed as an approach that provides insight into the workings of an Emergency Department. Simulation provides the ability to test 'what if' scenarios without actually incurring the real cost and risk associated with injecting change into the current critical real-world system (Jun, Jaconbson, & Swisher, 1999). For example, authors have used System Dynamics to represent and study issues in Emergency Department crowding. Lane, Monefedlt, and Rosenhead (2000) use system dynamic modeling to investigate the reduction in bed capacity to the waiting times of patients and Behr and Diaz (2010) use Systems Dynamics to model sensitivity of Emergency Department utilization.

Existing reviews of the use of modeling and simulation to better understand Emergency Departments show that much research has been focused on well-defined aspects or particular operations within or related to the Emergency Department (Jun, Jaconbson, & Swisher, 1999). Many papers have a focus on the causes and solutions for specific emergency Department issues (Gunal & Pidd, 2009)

At a higher level, modeling and simulating the Emergency Department as part of a larger healthcare system is also possible. The present study aims to provide the reader a holistic view of the various factors that affect the Emergency Department utilization.

System Dynamics allows for a systemic view of the variables interactions at an aggregated level. However, there are disadvantage with the approach as loss of effects of stochastic variation and resolution down to individual patient or condition level. Nevertheless, healthcare provisions cannot be understood by looking at factors in isolation. System Dynamics is the tool of choice since it offers the ability to produce technically representative models that are persuasive to stakeholders.

3.0 RESEARCH QUESTION

Policy makers and hospital managers are faced with central decisions in relation to incentivizing and expanding healthcare infrastructure and increasing capacities. When and how to expand the supply of ambulatory healthcare services are difficult since such decisions may involve large personnel and capital commitments as well as have the potential to alter the demand for services at other existing ambulatory care venues. These decisions must be made in the context of market forces that require a health facility to sustain its economic viability. Therefore, understanding the nonlinear dynamics inherent in the interaction among supply and demand variables is critical to healthcare management sustainability.

The central focus of this research is to provide an explication of the salient factors at play within the dynamic demand-supply tensions within which ambulatory care is provided within an Emergency Department. In addition, this research answers question related to the relative importance of such factors in reaching near demand-supply equilibrium and the cost implications of approaching such a solution.

4.0 RESEARCH APPROACH

This research proposes a Modeling and Simulation (M&S) approach based on System Dynamics to represent and simulate ambulatory ED utilization. The four basic steps that gird our approach are as follows: 1) represent the general flow of patients seeking ambulatory care, 2) represent the main venues that serve ambulatory patients with a particular focus on service provided by Emergency Departments, 3) represent the driving forces that animate a patient's willingness to seek services from a particular treatment venue, and 4) simulate the system and validate results through comparative analysis with national data. In the following sections, the authors describe the salient features of the model that include the venues, patient population, selection process, capacity model, and revenues model.

4.1 Ambulatory treatment venues

Ambulatory care can be broadly classified into two categories, namely office-based primary care offered by independent physician practitioners and hospital-based healthcare centers. For practical purposes, in this study we consolidate the subcategories that compose each of these as either primary care physician (PCP) or Emergency Department (ED) venues. The PCP will consider the combined data for the three categories of office-based primary care physicians while the ED will consider the combined data for the emergency departments and the outpatient department. The insurance status of treatment seekers also distinguishes PCP and ED venues. Patients having no insurance (self-pay, no charge, or charity) have a greater tendency to visit the ED (inclusive of the OPD) realtive the PCP. Pitts, Niska, Xu, and Bur (Pitts, Niska, Xu, & Burt, 2008) indicate that 17.4% are uninsured patients as a percentage of total patients.

4.2 Patient Population

In the presented model, the patient population at any given time is assumed to be a fraction of the total population. In addition, we assume that the patient population is aging and the cumulative demand for ambulatory care is increasing. This is supported by knowledge of the aging U.S. baby boomer population. Hence, the cumulative healthcare needs of the society have been increasing. The argument is found to be valid for the data between 1999-2006. This assumption is supported by Burt and Schappert, (2004), and Cherry, Hing, Woodwell, and Rechtsteiner (2008).

An important determinant of treatment seeking behavior from ambulatory care providers is insurance status (Behr, 2008). In this model, we explicitly consider this status as a major factor. In general, the patient population is split into the two categories of insured (Private, Medicare, Medicaid, State children's health insurance program) and uninsured (self-pay, no charge, or charity). According to the U.S. Census Bureau, 1999-2008, 84.2% of the population had some form of insurance with the remainder (15.8%) classified as uninsured. Utilization of healthcare venues between the insured and the uninsured is uneven. Ambulatory visits by insured patients are 93.69%. This parameter (93.69%) is used to divide the patient population into insured and uninsured categories.

4.3 Combined weight for selection of Emergency Department

When seeking ambulatory care, individuals tend to select a venue (supply) that closely matches their needs (demand). The venue selection is based on a set of core factors including access, capacity, waiting time to be served, and financial status. The weight of each factor in the individual decision calculus to seek services from one venue relative another may be derived from surveying the population that seeks ambulatory care (2008). There are many scales and methodologies for quantifying these values. In our case, for practical purposes, we selected ED access and PCP capacity as the main drivers for selecting a given venue. Calibration values at the start of simulation runs accounted for representing factors other than those indicated above.

4.4 Combined weight for selection for ER by uninsured patients

The choice of venue by uninsured patients is computed in similar fashion as those computed for insured patients. However, three additional factors are considered: tendency to defer treatment, insurance status, and level of patient acuity.

Within the model, we imbue uninsured patients with a greater tendency to visit an ED relative a PCP. The rational for this assumption is that the 'Tendency to defer treatment' and the 'insurance status' factors will reduce the chance of the patient visiting either the PCP or the ED. Uninsured patients have a greater tendency to defer medical treatment due to financial constraints relative the insured and, thus, the uninsured are more likely to self treat. This tendency has been adopted in the model via the 'Tendency to defer treatment' factor. This can be represented in our model by means of the factor 'Patient acuity level'. The uninsured patient has the tendency to defer treatment, which may eventually lead to a higher acuity. A person with a medical condition requiring immediate stabilization is more likely to seek services from an ED relative a PCP. Thus, the two factors, 'Tendency to defer treatment' and 'insurance status,' promote a tendency to avoid an ED visit whereas the 'Patient acuity level' would promote the choice of the ED instead of the PCP.

4.5 Capacity Submodel

The capacity is modeled in terms of the capability of the system to treat a certain number of patients. The capacity is increased or reduced depending on the difference between available and target capacity. The available capacity and the number of patients visiting the facility is classically defined the 'Beds per person.' This represents units of capacity available per patient. This parameter influences the estimated waiting time at the facility 'ED waiting time factor' and is modeled as a regression with 'beds per patient.' As indicated before, the 'ED waiting time' factor is one of the aspects that contribute to the 'Combined weight for selection of ED'. The same model is implemented in case of PCP. The below Figure 1 below displays this model.

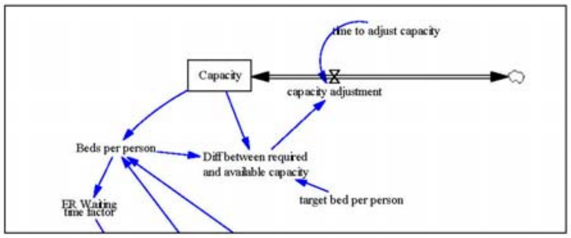


Figure 1 – Capacity submodel

4.6 Revenue submodel

The revenue submodel estimates the overall revenue generated by a particular venue as function of income per patient ('price') and the number of patients visiting the venue. The 'unpaid healthcare costs' are modeled as the 'price' and the number of uninsured patient visits. Figure 2 and 3 below show this simple revenue submodel and the accumulative stock for the 'unpaid healthcare costs.' The 'unpaid healthcare costs' is exponentially impacted by the tendency to defer treatment, which has been known to deteriorate and complicate the medical condition of the patient. A worsening of medical condition leads to higher costs.

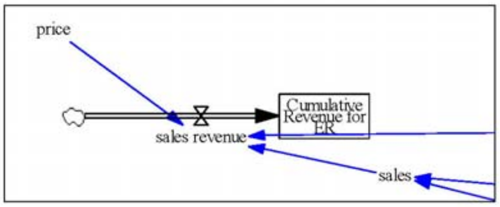


Figure 2 – Revenue model for healthcare venue

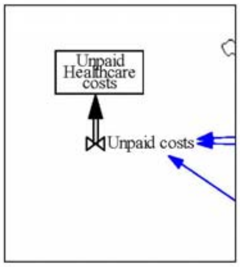


Figure 3 – Unpaid healthcare costs

5.0 RESULTS

The usefulness of the presented simulation model is reflected in its close approximation to the real-world system. That is, a model whose salient variables function in a fashion similar to patterns apparent in the real-world is often useful. As such, a classic method to measure the quality of the performance of the simulation model is to compare its output values with actual available statistics. If the real-world system has been properly modeled, then the differences between the two systems measures of performance ought to be minimal. The measures of performance selected to quantify the performance of the modeled system includes: Total Number of Visits to ED (divided between insured and uninsured patients), the Total Number of Visits to PCP (divided between insured and uninsured patients). The results are displayed in Table 1 below as a comparison of simulated and actual measures of performance.

Table 1 – Comparison of simulated and actual parameters

Year -2006	**Simulated**	**Actual**	**% deviation**
Total Number of Visits to ED	*217,642,000*	*221,399,000*	*-1.73%*
Number of Visits to ER - Insured patients	186,065,320	190,234,550	-2.24%
Number of Visits to ER - Uninsured patients	31,576,770	31,164,450	1.31%
Total Number of Visits to PCP	*904,368,000*	*901,954,000*	*0.27%*
Number of Visits to PCP -Insured patients	863,874,700	862,268,024	0.19%
Number of Visits to PCP -Uninsured patients	40,492,735	39,685,976	1.99%

It can be observed that the simulated and actual values closely match. This demonstrates that the level to which this model is calibrated to the actual system and also provides initial validation for the overall construction of the model.

6.0 DISCUSSION AND CONCLUSIONS

System Dynamics is a simulation approach that can be used to capture a holistic perspective of the ambulatory care system. In this paper, we developed a SD model that replicate the behavior of the ambulatory care system. The model considered both the supply and demand sides of the system. Further, it considers specific factors that influence individuals' decisions in selecting ambulatory care venues. Venues were categorized as either EDs for hospital-based care centers or PCPs for office-based primary care physicians. ED access and PCP capacity were selected as major drivers that conditioned the venue selection. Tendency to defer treatment, insurance status, and patient acuity level also are critical aspects that influence the performance of the model, although other components of the model, namely capacity and revenues submodels, were not fully analyzed in this paper.

The simulated model was executed for replicating the period from 1999 through 2006. Empirical results demonstrate that the model perform well. To measure the performance of the simulated data a comparisons were made relative real-world data. Deviation of the simulated data from real data was approximately 1.29%. Measures of performance included: total patients visits to EDs and PCPs for insured and uninsured.

Potential application of this simulation are manifold, including the following: 1) investigating the mix of intervention techniques that divert targeted patients to alternative venues, 2) the capacity and financial consequences of expanding or downsizing different venues, 3) the effects of delaying treatment, 4) providing ambulatory care to certain segment of the population, and 5) the consequences of congestion on demand and supply factors.

The simulation model demonstrates a handling of the complexities associated with an ambulatory care system. Managerial and policy-making decision environments systems require effective tools to support the decision process. Simulation-based decision support systems that embrace these technologies can assertively center on efforts and resource allocation that produce demonstrable sustainable solutions.

7.0 REFERENCES

[1] Aday, L., & Andersen, R. (1974). *A framework for the study of access to medical care.* Health Services Research.

[2] Behr, J. (2008). *Drivers of Emergency Department Utilization: A Systematic Study of the Individual's Decision Calculus to Select Healthcare Services.* Sentara Health Foundation. Norfolk, VA: Unpublished.

[3] Behr, J., & Diaz, R. (2010). A system dynamics approach to modeling the sesnititvity of inappropiate emerency department utilization. *Third International Conference on Social Computing, behavioral modeling, and prediction. SBP 2010.* Bethesda, MD: Advances in Social Computing - Springer.

[4] Burt, C. W., & Schappert, S. M. (2004). Ambulatory care visits to physician offices, hospital outpatient departments, and emergency departments: United States, 1999– 2000, National Center for Health Statistics. Vital Health Stat. *13* (157).

[5] Carr, B., Kaye, A., Wiebe, D., Gracias, V., Schwab, C., & Reilly, P. (2007). Emergency department length of stay: a major risk factor for pneumonia in intubated blunt trauma patients. *63* (9-12).

[6] Cherry, D. K., Hing, E., Woodwell, D. A., & Rechtsteiner, E. A. (2008). *National Ambulatory Medical Care Survey: 2006 Summary, National Health Statistics Reports.* Centers for Disease Control and Prevention, U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES.

[7] Grey, B., Gusmano, M., & Collins, S. (2003). *AHCPR and the changing politics of health services research.* Health Services Research.

[8] Gunal, M., & Pidd, M. (2009). Discrete Event Simulation for Performance Modeling in Healthcare. *Unpublished,Lanchaster University Management School, Working Paper* .

[9] Jun, J. B., Jaconbson, S. H., & Swisher, J. R. (1999). Application of discrete-event simulation in health care clinics: A survey. *Journal of the Operational Research Society , 50* (2), 109-123.

[10] Lane, D. C., Monefedlt, C., & Rosenhead, J. V. (2000). Looking in the Wrong Place for Healthcare Improvements: A System Dynamics Study of an Accident and Emergency Department. *The Journal of the Operational Research Society , 51* (5), 518-531.

[11] Morgan, M. W., Zamora, N. E., & F., H. M. (2007). An Inconvenient Truth: A Sustainable Healthcare System Requires Chronic Disease Prevention and Management Transformation. *7* (4).

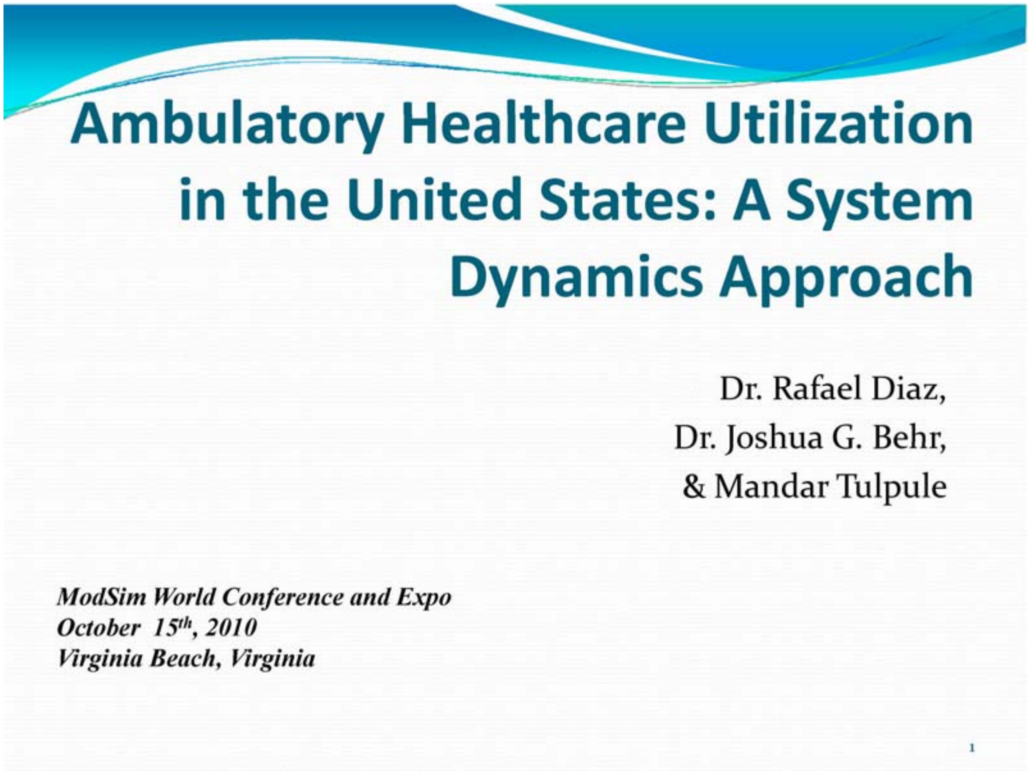
[12] Pitts, S. R., Niska, R. W., Xu, J., & Burt, C. W. (2008). *National Hospital Ambulatory Medical Care Survey: 2006 Emergency Department Summary, National Health Statistics Reports.* Centers for Disease Control and Prevention, U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES.

[13] U.S. Census Bureau. (2009, December). *Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2000 to July 1, 2009 (NST-EST2009-01).* Retrieved July 20, 2010, from census.gov: http://www.census.gov/popest/states/NST-ann-est.html

[14] U.S. Census Bureau. (1999-2008). *Health Insurance Historical Tables, Health Insurance Coverage Status and Type of Coverage by Nativity: 1999 to 2008.*

Retrieved July 20, 2010, from census.gov: http://www.census.gov/hhes/www/hlthins/data/historical/index.html

[15] Williams, D. (1994). *The concept of race in health services research..* Health Services Research.

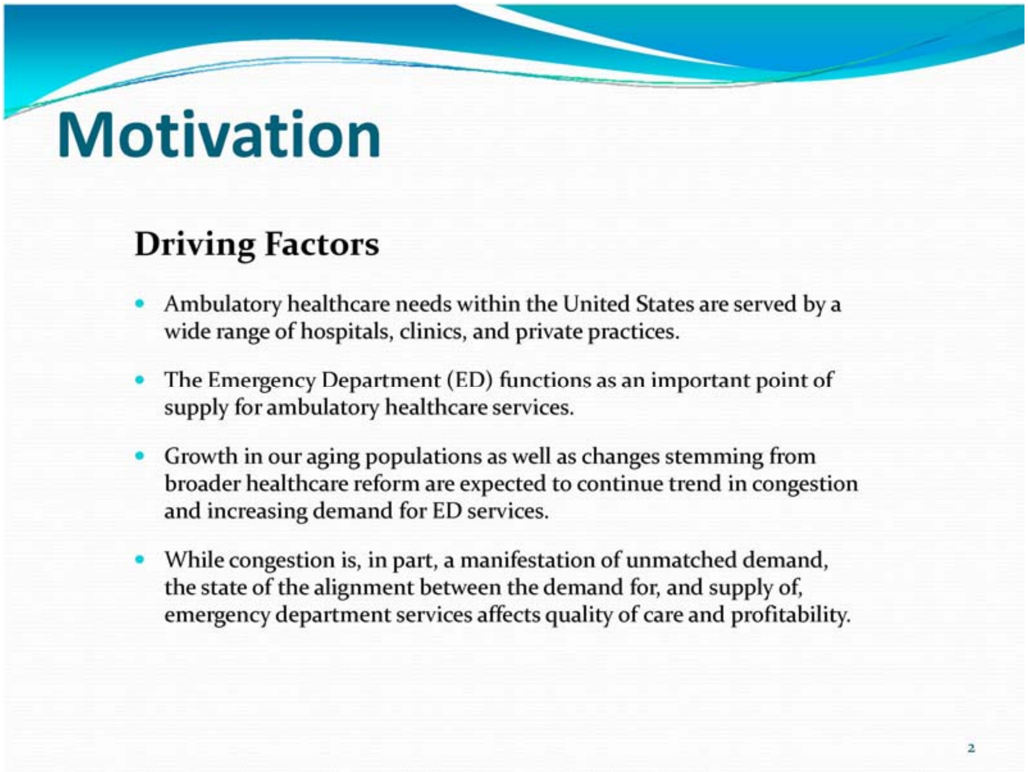


Ambulatory Healthcare Utilization in the United States: A System Dynamics Approach

Dr. Rafael Diaz,
Dr. Joshua G. Behr,
& Mandar Tulpule

ModSim World Conference and Expo
October 15th, 2010
Virginia Beach, Virginia

1

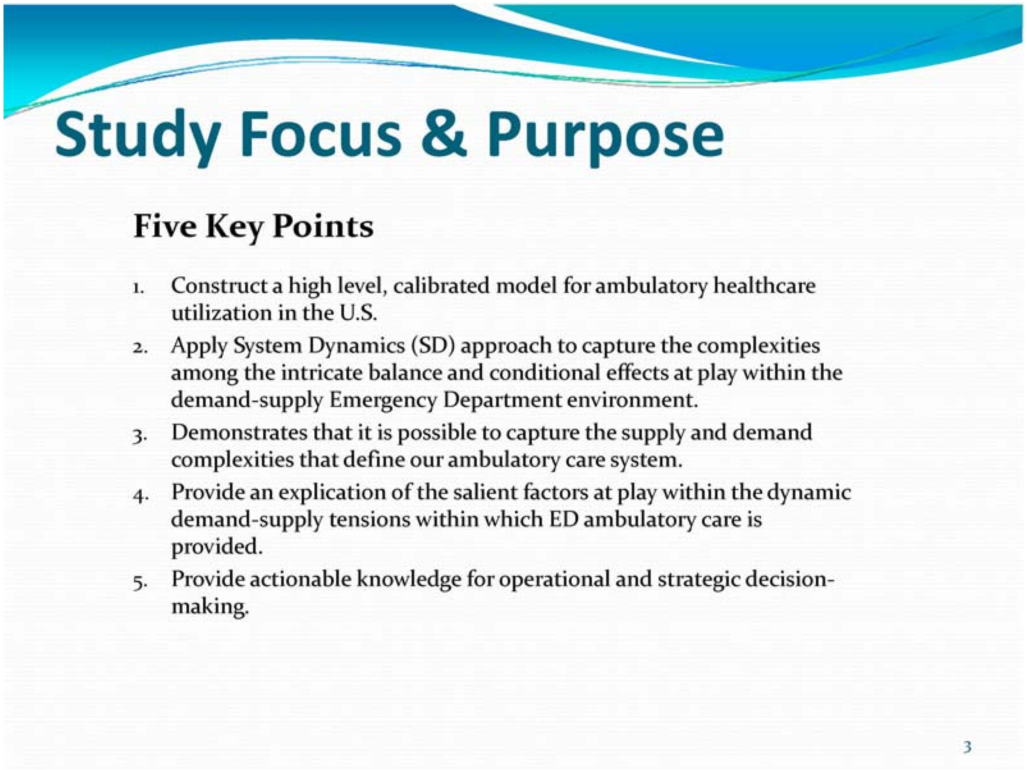


Motivation

Driving Factors

- Ambulatory healthcare needs within the United States are served by a wide range of hospitals, clinics, and private practices.
- The Emergency Department (ED) functions as an important point of supply for ambulatory healthcare services.
- Growth in our aging populations as well as changes stemming from broader healthcare reform are expected to continue trend in congestion and increasing demand for ED services.
- While congestion is, in part, a manifestation of unmatched demand, the state of the alignment between the demand for, and supply of, emergency department services affects quality of care and profitability.

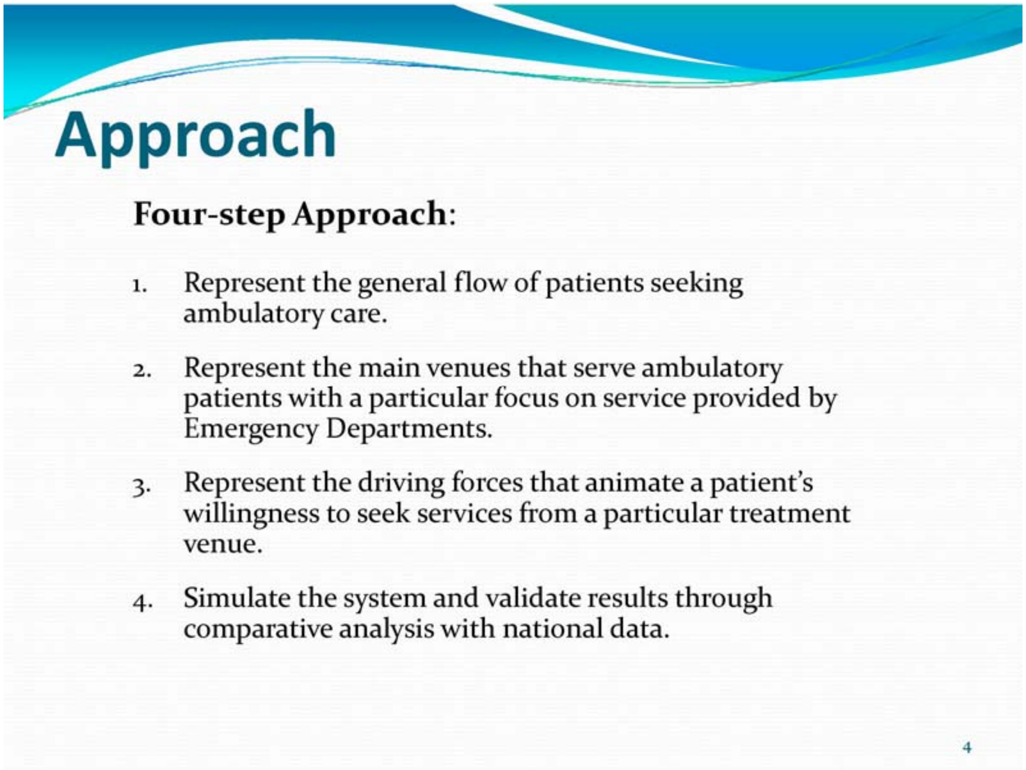
2



Study Focus & Purpose

Five Key Points

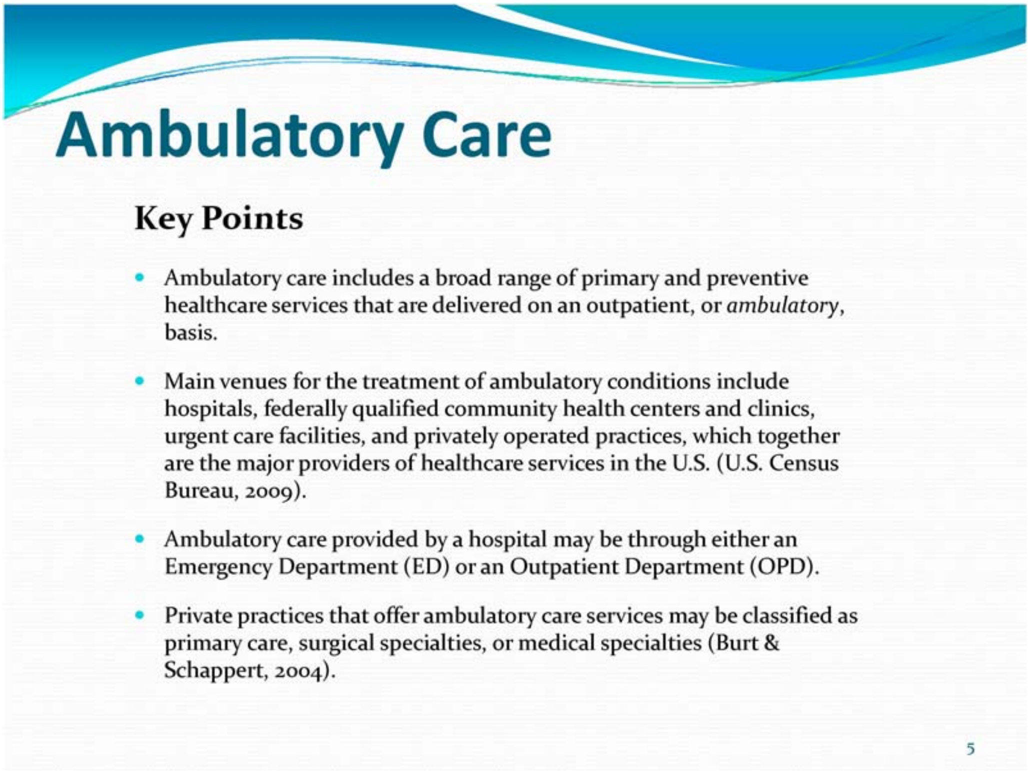
1. Construct a high level, calibrated model for ambulatory healthcare utilization in the U.S.
2. Apply System Dynamics (SD) approach to capture the complexities among the intricate balance and conditional effects at play within the demand-supply Emergency Department environment.
3. Demonstrates that it is possible to capture the supply and demand complexities that define our ambulatory care system.
4. Provide an explication of the salient factors at play within the dynamic demand-supply tensions within which ED ambulatory care is provided.
5. Provide actionable knowledge for operational and strategic decision-making.



Approach

Four-step Approach:

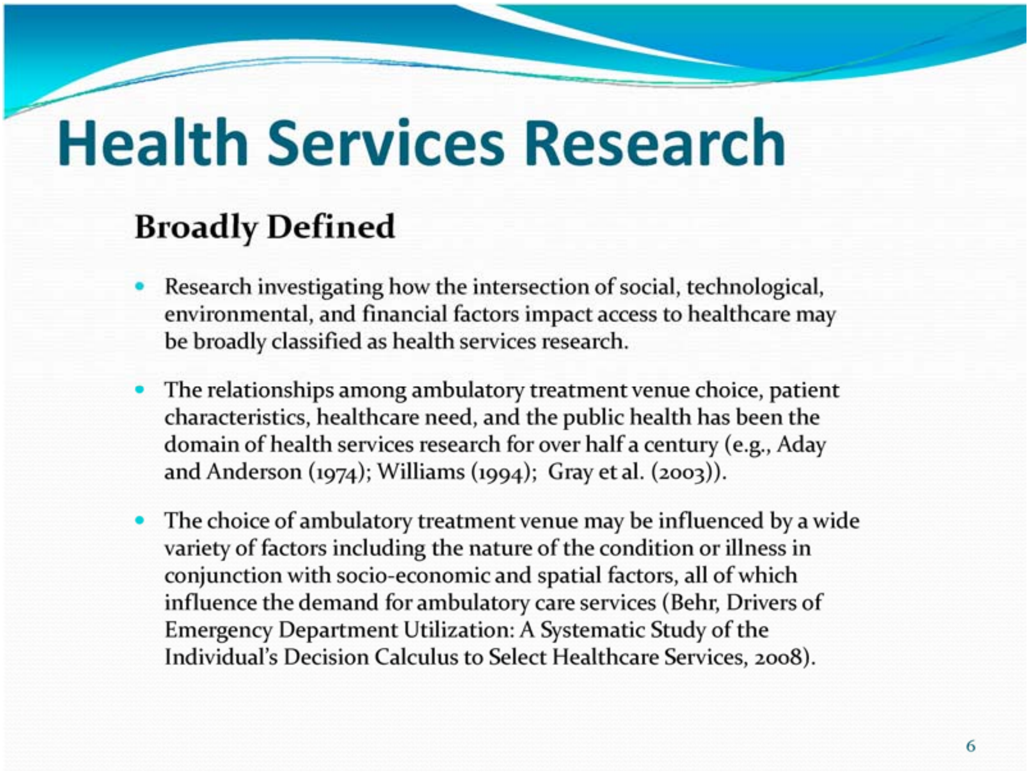
1. Represent the general flow of patients seeking ambulatory care.
2. Represent the main venues that serve ambulatory patients with a particular focus on service provided by Emergency Departments.
3. Represent the driving forces that animate a patient's willingness to seek services from a particular treatment venue.
4. Simulate the system and validate results through comparative analysis with national data.



Ambulatory Care

Key Points

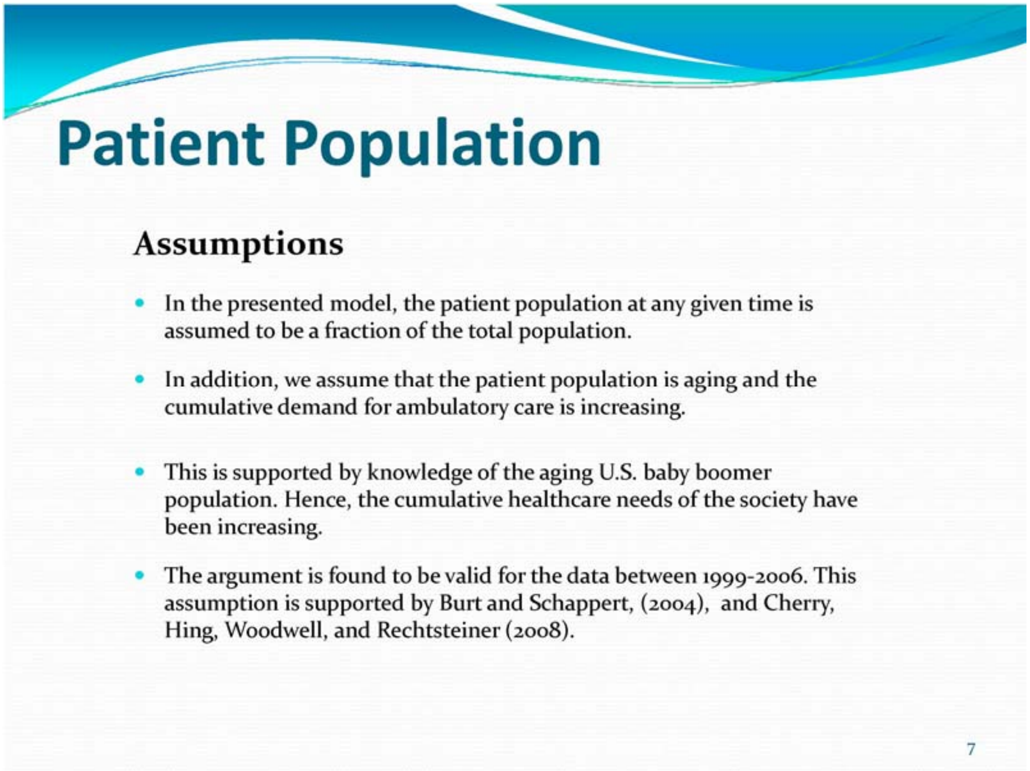
- Ambulatory care includes a broad range of primary and preventive healthcare services that are delivered on an outpatient, or *ambulatory*, basis.
- Main venues for the treatment of ambulatory conditions include hospitals, federally qualified community health centers and clinics, urgent care facilities, and privately operated practices, which together are the major providers of healthcare services in the U.S. (U.S. Census Bureau, 2009).
- Ambulatory care provided by a hospital may be through either an Emergency Department (ED) or an Outpatient Department (OPD).
- Private practices that offer ambulatory care services may be classified as primary care, surgical specialties, or medical specialties (Burt & Schappert, 2004).



Health Services Research

Broadly Defined

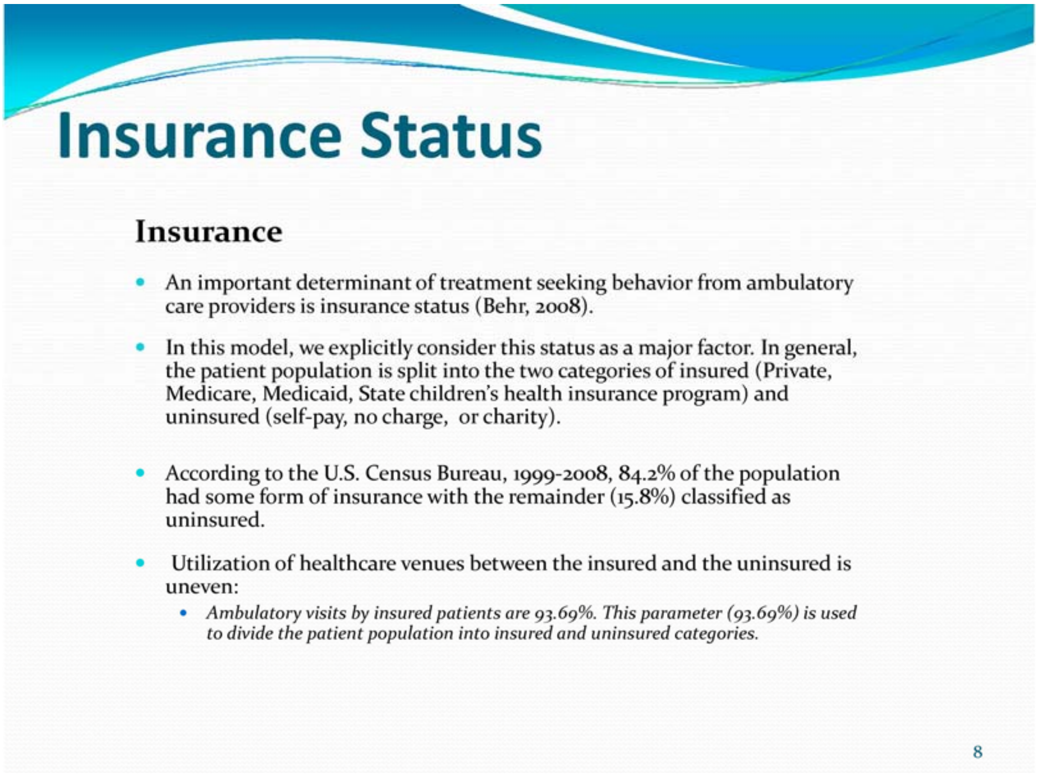
- Research investigating how the intersection of social, technological, environmental, and financial factors impact access to healthcare may be broadly classified as health services research.
- The relationships among ambulatory treatment venue choice, patient characteristics, healthcare need, and the public health has been the domain of health services research for over half a century (e.g., Aday and Anderson (1974); Williams (1994); Gray et al. (2003)).
- The choice of ambulatory treatment venue may be influenced by a wide variety of factors including the nature of the condition or illness in conjunction with socio-economic and spatial factors, all of which influence the demand for ambulatory care services (Behr, Drivers of Emergency Department Utilization: A Systematic Study of the Individual's Decision Calculus to Select Healthcare Services, 2008).



Patient Population

Assumptions

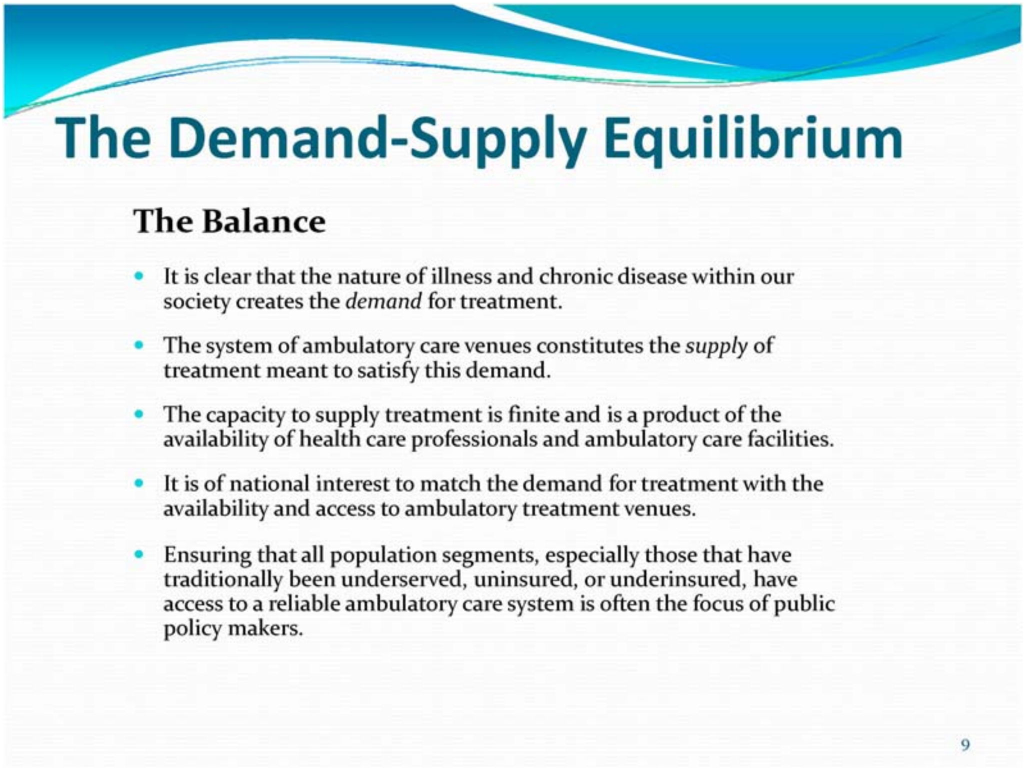
- In the presented model, the patient population at any given time is assumed to be a fraction of the total population.
- In addition, we assume that the patient population is aging and the cumulative demand for ambulatory care is increasing.
- This is supported by knowledge of the aging U.S. baby boomer population. Hence, the cumulative healthcare needs of the society have been increasing.
- The argument is found to be valid for the data between 1999-2006. This assumption is supported by Burt and Schappert, (2004), and Cherry, Hing, Woodwell, and Rechtsteiner (2008).



Insurance Status

Insurance

- An important determinant of treatment seeking behavior from ambulatory care providers is insurance status (Behr, 2008).
- In this model, we explicitly consider this status as a major factor. In general, the patient population is split into the two categories of insured (Private, Medicare, Medicaid, State children's health insurance program) and uninsured (self-pay, no charge, or charity).
- According to the U.S. Census Bureau, 1999-2008, 84.2% of the population had some form of insurance with the remainder (15.8%) classified as uninsured.
- Utilization of healthcare venues between the insured and the uninsured is uneven:
  - *Ambulatory visits by insured patients are 93.69%. This parameter (93.69%) is used to divide the patient population into insured and uninsured categories.*

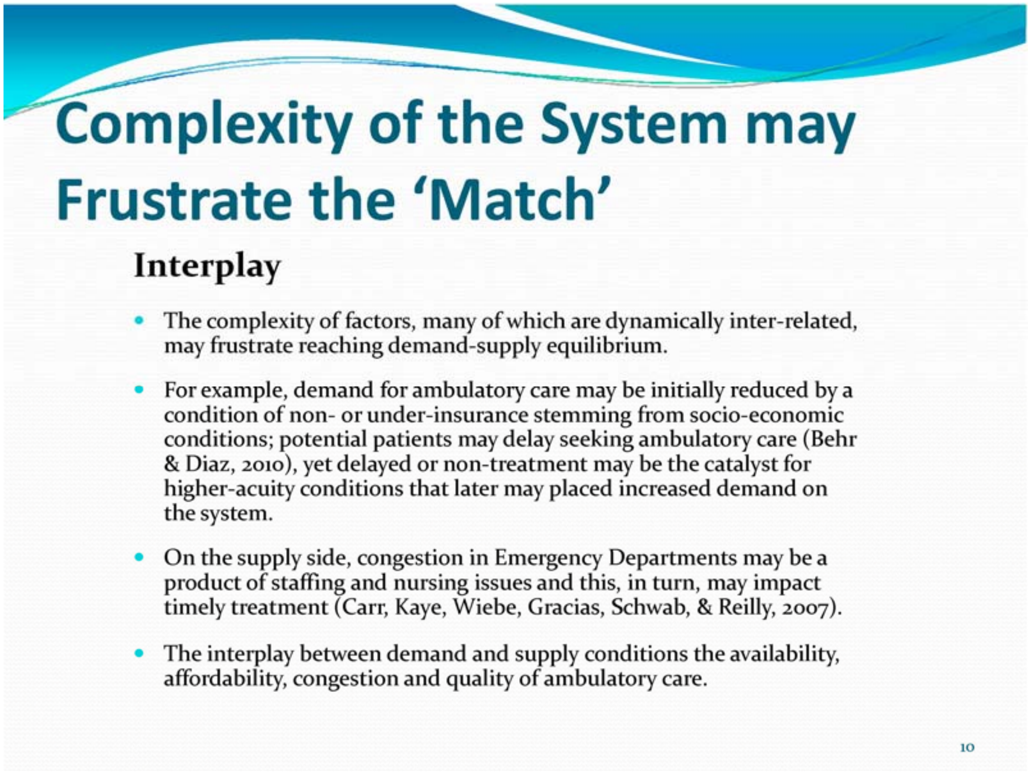


The Demand-Supply Equilibrium

The Balance

- It is clear that the nature of illness and chronic disease within our society creates the *demand* for treatment.
- The system of ambulatory care venues constitutes the *supply* of treatment meant to satisfy this demand.
- The capacity to supply treatment is finite and is a product of the availability of health care professionals and ambulatory care facilities.
- It is of national interest to match the demand for treatment with the availability and access to ambulatory treatment venues.
- Ensuring that all population segments, especially those that have traditionally been underserved, uninsured, or underinsured, have access to a reliable ambulatory care system is often the focus of public policy makers.

9

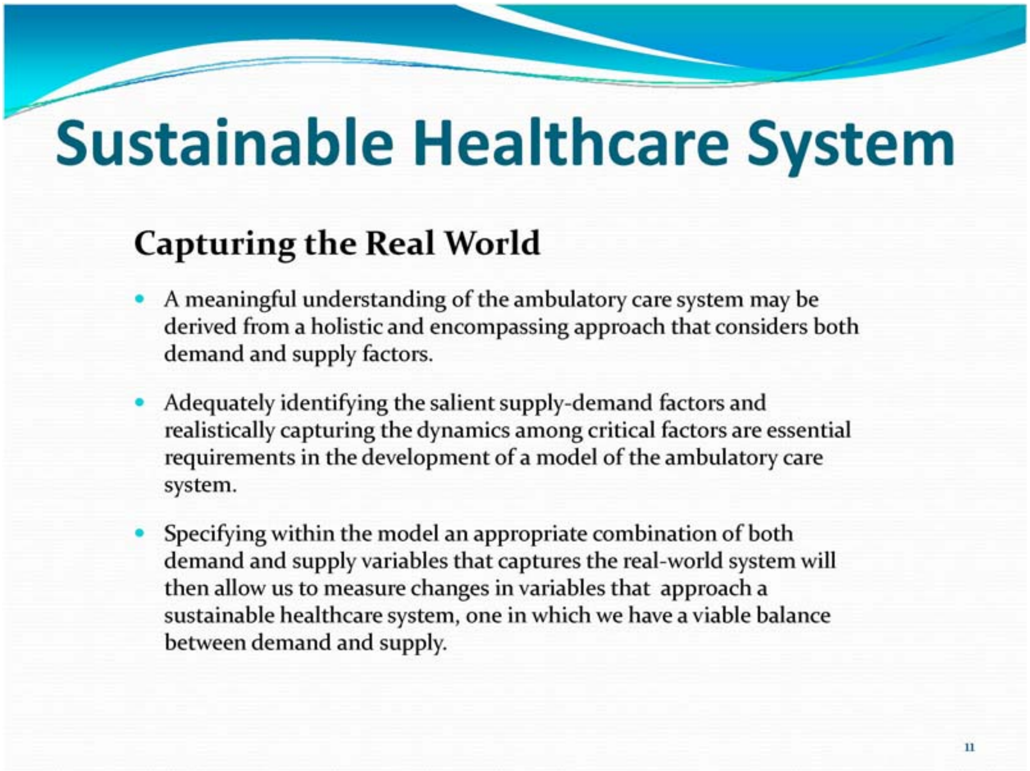


Complexity of the System may Frustrate the 'Match'

Interplay

- The complexity of factors, many of which are dynamically inter-related, may frustrate reaching demand-supply equilibrium.
- For example, demand for ambulatory care may be initially reduced by a condition of non- or under-insurance stemming from socio-economic conditions; potential patients may delay seeking ambulatory care (Behr & Diaz, 2010), yet delayed or non-treatment may be the catalyst for higher-acuity conditions that later may placed increased demand on the system.
- On the supply side, congestion in Emergency Departments may be a product of staffing and nursing issues and this, in turn, may impact timely treatment (Carr, Kaye, Wiebe, Gracias, Schwab, & Reilly, 2007).
- The interplay between demand and supply conditions the availability, affordability, congestion and quality of ambulatory care.

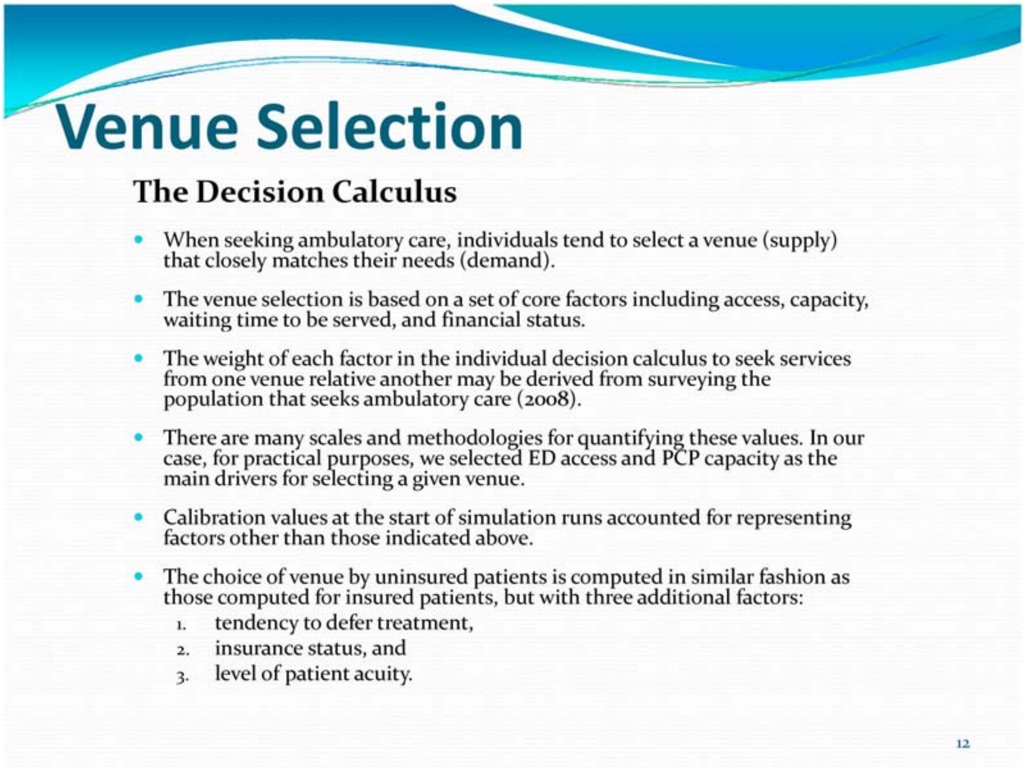
10



Sustainable Healthcare System

Capturing the Real World

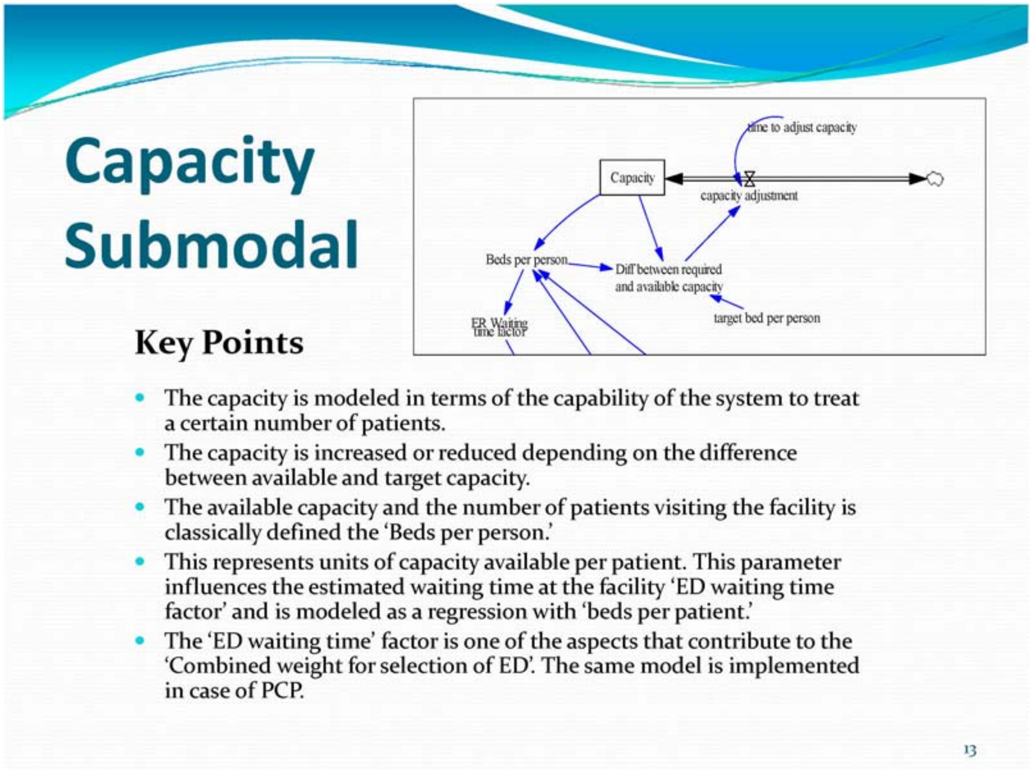
- A meaningful understanding of the ambulatory care system may be derived from a holistic and encompassing approach that considers both demand and supply factors.
- Adequately identifying the salient supply-demand factors and realistically capturing the dynamics among critical factors are essential requirements in the development of a model of the ambulatory care system.
- Specifying within the model an appropriate combination of both demand and supply variables that captures the real-world system will then allow us to measure changes in variables that approach a sustainable healthcare system, one in which we have a viable balance between demand and supply.



Venue Selection

The Decision Calculus

- When seeking ambulatory care, individuals tend to select a venue (supply) that closely matches their needs (demand).
- The venue selection is based on a set of core factors including access, capacity, waiting time to be served, and financial status.
- The weight of each factor in the individual decision calculus to seek services from one venue relative another may be derived from surveying the population that seeks ambulatory care (2008).
- There are many scales and methodologies for quantifying these values. In our case, for practical purposes, we selected ED access and PCP capacity as the main drivers for selecting a given venue.
- Calibration values at the start of simulation runs accounted for representing factors other than those indicated above.
- The choice of venue by uninsured patients is computed in similar fashion as those computed for insured patients, but with three additional factors:
  1. tendency to defer treatment,
  2. insurance status, and
  3. level of patient acuity.



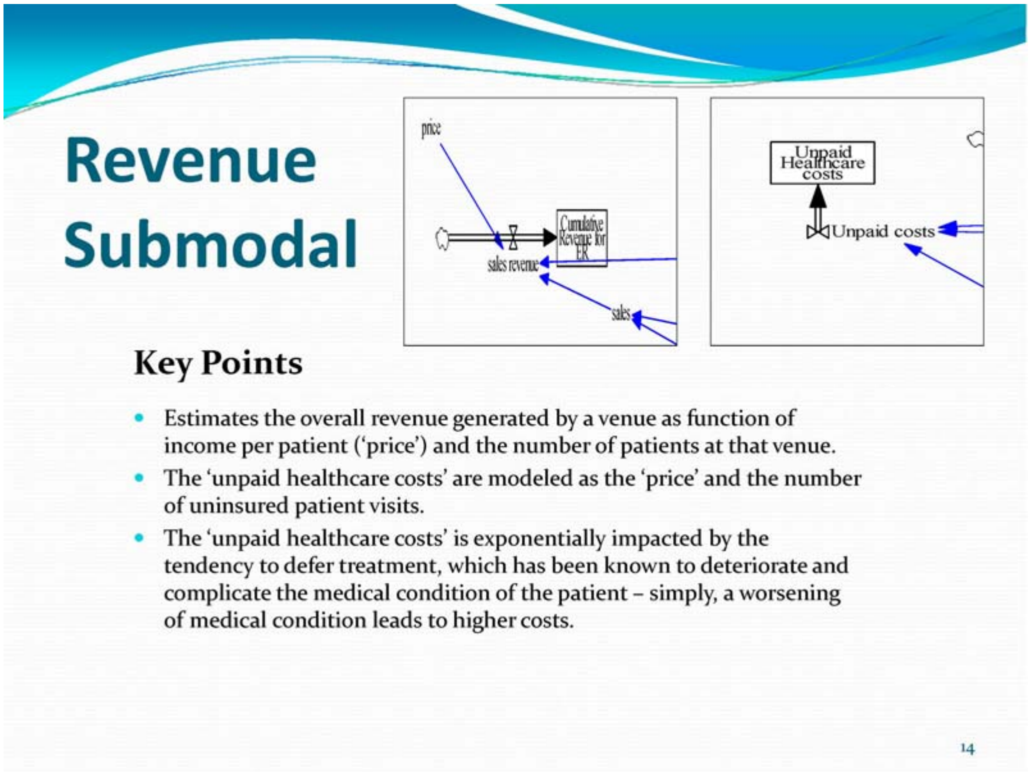
Capacity Submodal

time to adjust capacity
Capacity
capacity adjustment
Beds per person
Diff between required and available capacity
target bed per person
ER Waiting time factor

Key Points

- The capacity is modeled in terms of the capability of the system to treat a certain number of patients.
- The capacity is increased or reduced depending on the difference between available and target capacity.
- The available capacity and the number of patients visiting the facility is classically defined the 'Beds per person.'
- This represents units of capacity available per patient. This parameter influences the estimated waiting time at the facility 'ED waiting time factor' and is modeled as a regression with 'beds per patient.'
- The 'ED waiting time' factor is one of the aspects that contribute to the 'Combined weight for selection of ED'. The same model is implemented in case of PCP.

13



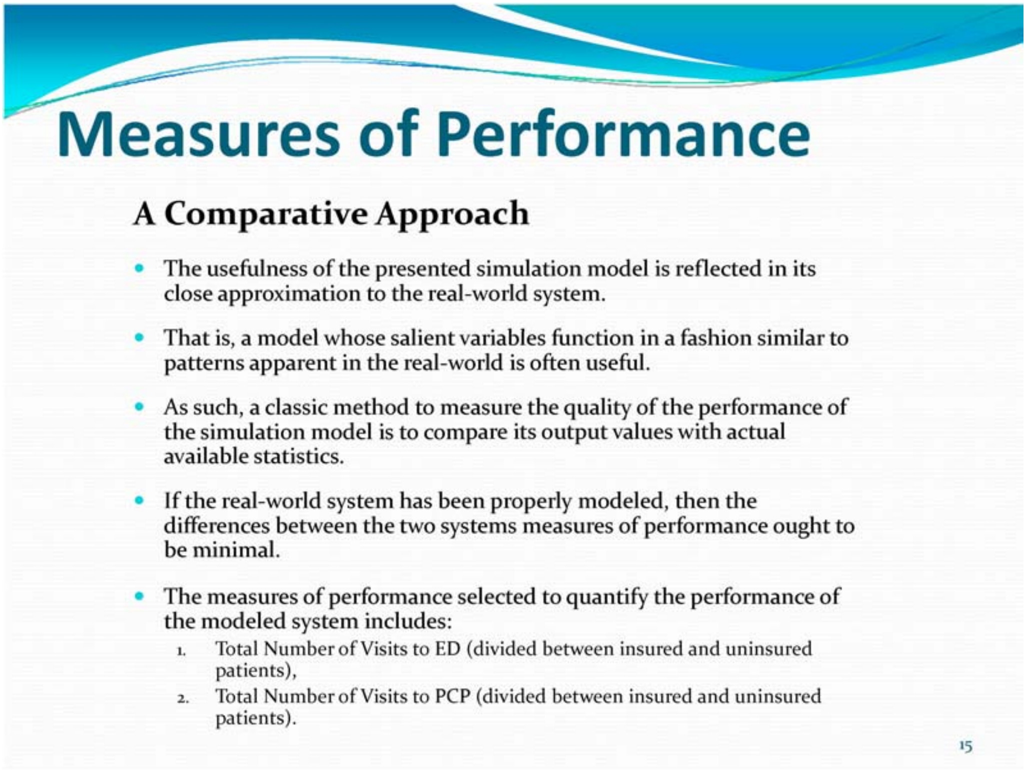
Revenue Submodal

price
Cumulative Revenue for ER
sales revenue
sales
Unpaid Healthcare costs
Unpaid costs

Key Points

- Estimates the overall revenue generated by a venue as function of income per patient ('price') and the number of patients at that venue.
- The 'unpaid healthcare costs' are modeled as the 'price' and the number of uninsured patient visits.
- The 'unpaid healthcare costs' is exponentially impacted by the tendency to defer treatment, which has been known to deteriorate and complicate the medical condition of the patient – simply, a worsening of medical condition leads to higher costs.

14

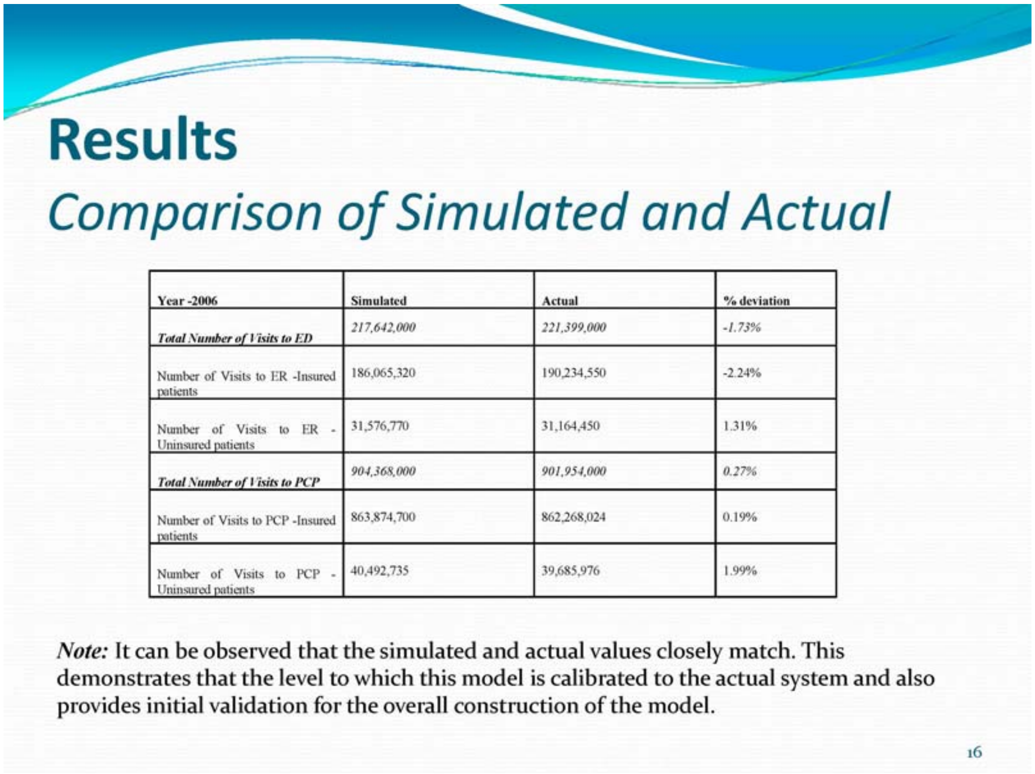


Measures of Performance

A Comparative Approach

- The usefulness of the presented simulation model is reflected in its close approximation to the real-world system.
- That is, a model whose salient variables function in a fashion similar to patterns apparent in the real-world is often useful.
- As such, a classic method to measure the quality of the performance of the simulation model is to compare its output values with actual available statistics.
- If the real-world system has been properly modeled, then the differences between the two systems measures of performance ought to be minimal.
- The measures of performance selected to quantify the performance of the modeled system includes:
  1. Total Number of Visits to ED (divided between insured and uninsured patients),
  2. Total Number of Visits to PCP (divided between insured and uninsured patients).

15



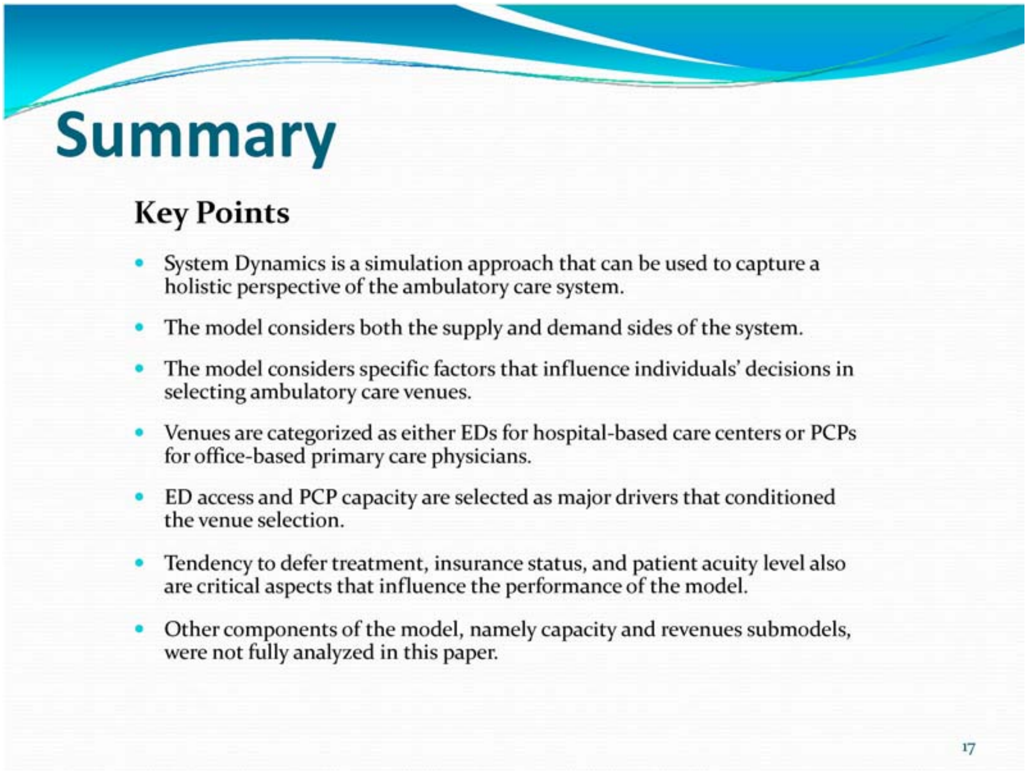
Results

Comparison of Simulated and Actual

Year -2006	**Simulated**	**Actual**	**% deviation**
Total Number of Visits to ED	*217,642,000*	*221,399,000*	*-1.73%*
Number of Visits to ER -Insured patients	186,065,320	190,234,550	-2.24%
Number of Visits to ER - Uninsured patients	31,576,770	31,164,450	1.31%
Total Number of Visits to PCP	*904,368,000*	*901,954,000*	*0.27%*
Number of Visits to PCP -Insured patients	863,874,700	862,268,024	0.19%
Number of Visits to PCP - Uninsured patients	40,492,735	39,685,976	1.99%

Note: It can be observed that the simulated and actual values closely match. This demonstrates that the level to which this model is calibrated to the actual system and also provides initial validation for the overall construction of the model.

16

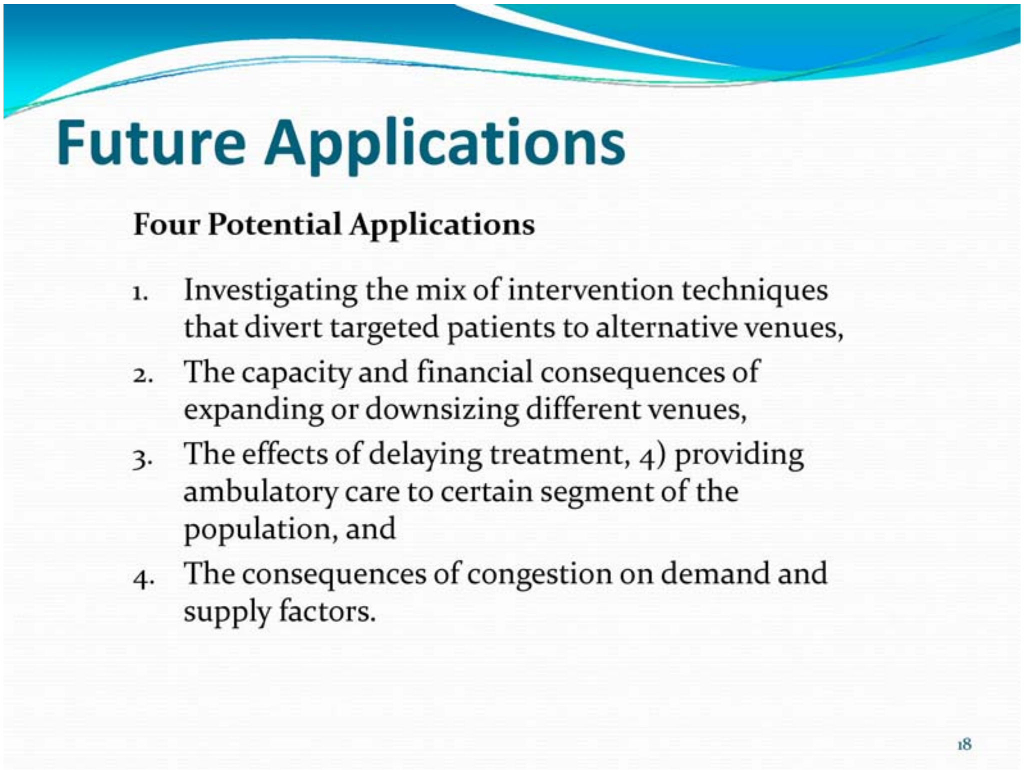


Summary

Key Points

- System Dynamics is a simulation approach that can be used to capture a holistic perspective of the ambulatory care system.
- The model considers both the supply and demand sides of the system.
- The model considers specific factors that influence individuals' decisions in selecting ambulatory care venues.
- Venues are categorized as either EDs for hospital-based care centers or PCPs for office-based primary care physicians.
- ED access and PCP capacity are selected as major drivers that conditioned the venue selection.
- Tendency to defer treatment, insurance status, and patient acuity level also are critical aspects that influence the performance of the model.
- Other components of the model, namely capacity and revenues submodels, were not fully analyzed in this paper.

17

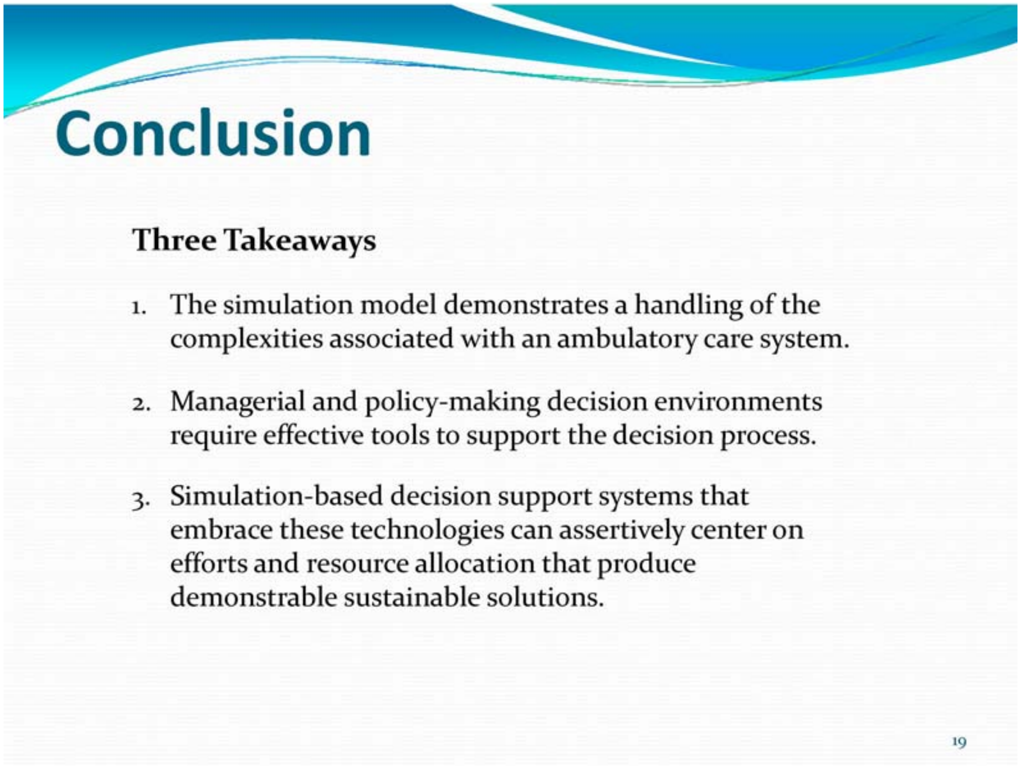


Future Applications

Four Potential Applications

1. Investigating the mix of intervention techniques that divert targeted patients to alternative venues,
2. The capacity and financial consequences of expanding or downsizing different venues,
3. The effects of delaying treatment, 4) providing ambulatory care to certain segment of the population, and
4. The consequences of congestion on demand and supply factors.

18



Conclusion

Three Takeaways

1. The simulation model demonstrates a handling of the complexities associated with an ambulatory care system.
2. Managerial and policy-making decision environments require effective tools to support the decision process.
3. Simulation-based decision support systems that embrace these technologies can assertively center on efforts and resource allocation that produce demonstrable sustainable solutions.

19

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* 01- 03 - 2011	**2. REPORT TYPE** Conference Publication	**3. DATES COVERED** *(From - To)*

4. TITLE AND SUBTITLE Selected Papers and Presentations Presented at MODSIM World 2010 Conference & Expo	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Pinelli, Thomas E. (Editor)	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER 736466.01.04.07.08
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) NASA Langley Research Center Hampton, VA 23681-2199	**8. PERFORMING ORGANIZATION REPORT NUMBER** L-20000
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) National Aeronautics and Space Administration Washington, DC 20546-0001	**10. SPONSOR/MONITOR'S ACRONYM(S)** NASA
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) NASA/CP-2011-217069/Part 1

12. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified - Unlimited
Subject Category 66 System Analysis and Operations Research
Availability: NASA CASI (443) 757-5802

13. SUPPLEMENTARY NOTES
Thomas E. Pinelli, NASA Langley Research Center, Strategic Relationships Office; thomas.e.pinelli@nasa.gov;
Phone: (757) 864-2491

14. ABSTRACT

MODSIM World 2010 was held in Hampton, Virginia, October 13-15, 2010. The theme of the 2010 conference & expo was "21st Century Decision-Making: The Art of Modeling& Simulation". The conference program consisted of seven technical tracks - Defense, Engineering and Science, Health & Medicine, Homeland Security & First Responders, The Human Dimension, K-20 STEM Education, and Serious Games & Virtual Worlds. Selected papers and presentations from MODSIM World 2010 Conference & Expo are contained in this NASA Conference Publication (CP). Section 8.0 of this CP contains papers from MODSIM World 2009 Conference & Expo that were unavailable at the time of publication of NASA/CP-2010-216205 Selected Papers Presented at MODSIM World 2009 Conference and Expo, March 2010.

15. SUBJECT TERMS
Modeling and simulation; Agent-based modeling; Game theory; Discrete event simulation; Cognitive modeling; Visualization

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			STI Help Desk (email: help@sti.nasa.gov)
					19b. TELEPHONE NUMBER ***(Include area code)***
U	U	U	UU	477	(443) 757-5802